A Companion to David Lewis

Edited by

Barry Loewer and Jonathan Schaffer

WILEY Blackwell

This edition first published 2015
© 2015 John Wiley & Sons, Inc.

Registered Office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Offices
350 Main Street, Malden, MA 02148-5020, USA
9600 Garsington Road, Oxford, OX4 2DQ, UK
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, for customer services, and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell.

The right of Barry Loewer and Jonathan Schaffer to be identified as the authors of the editorial material in this work has been asserted in accordance with the UK Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty: While the publisher and authors have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. It is sold on the understanding that the publisher is not engaged in rendering professional services and neither the publisher nor the author shall be liable for damages arising herefrom. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Library of Congress Cataloging-in-Publication Data
Cataloging-in-Publication Data has been applied for.

C 9781118388181

A catalogue record for this book is available from the British Library.

Cover image: Princeton University, Office of Communications, photo: R. Mathews

Set in 9.5/11.5 pt PhotinaMTStd by Toppan Best-set Premedia Limited

1 2015

Table of Contents

Notes on Contributors ix

Part I Biography and New Work **1**

1 Intellectual Biography of David Lewis (1941–2001): Early Influences 3
Stephanie R. Lewis

2 Counterparts of States of Affairs 15
David Lewis

3 Reply to Dana Scott, "Is There Life on Possible Worlds?" 18
David Lewis

Part II Methodology and Context **23**

4 Lewis's Philosophical Method 25
Daniel Nolan

5 On Metaphysical Analysis 40
David Braddon-Mitchell and Kristie Miller

6 A Lewisian History of Philosophy 60
Robert Pasnau

7 David Lewis's Place in Analytic Philosophy 80
Scott Soames

Part III Metaphysics and Science **99**

8 Humean Supervenience 101
Brian Weatherson

9 No Work for a Theory of Universals 116
M. Eddon and C.J.G. Meacham

10 Hume's Dictum and Metaphysical Modality: Lewis's Combinatorialism 138
Jessica Wilson

11 Truthmaking: With and Without Counterpart Theory 159
Phillip Bricker

12 How to Be Humean 188
Jenann Ismael

13 Where (in Logical Space) Is God? 206
Stephanie R. Lewis

14 *De Re* Modality, Essentialism, and Lewis's Humeanism 220
Helen Beebee and Fraser MacBride

15 David Lewis on Persistence 237
Katherine Hawley

16 "Perfectly Understood, Unproblematic, and Certain": Lewis on Mereology 250
Karen Bennett

17 Humean Reductionism about Laws of Nature 262
Ned Hall

18 Why Lewisians Should Love Deterministic Chance 278
Rachael Briggs

19 Lewis on Causation 295
Christopher Hitchcock

Part IV Language and Logic 313

20 David Lewis on Convention 315
Ernie Lepore and Matthew Stone

21 Asking What a Meaning *Does*: David Lewis's Contributions to Semantics 328
Barbara H. Partee

22 Accommodation in a Language Game 345
Craige Roberts

23 Lewis on Reference and Eligibility 367
J.R.G. Williams

24 On the Nature of Certain Philosophical Entities: Set Theoretic Constructionalism in the Metaphysics of David Lewis 382
Gideon Rosen

25 Primitive Self-Ascription: Lewis on the *De Se* 399
Richard Holton

26 Counterfactuals and Humean Reduction 411
Robert Stalnaker

27 On the Plurality of Lewis's Triviality Results 425
Alan Hájek

28 Decision Theory after Lewis 446
John Collins

29 Lewis on Mereology and Set Theory 459
John P. Burgess

Part V Epistemology and Mind 471

30 Lewis on Knowledge Ascriptions 473
Jonathan Schaffer

31 Humility and Coexistence in Kant and Lewis: Two Modal Themes, with Variations 491
Rae Langton

32 Analytic Functionalism 504
Wolfgang Schwarz

33 Lewis on Materialism and Experience 519
Daniel Stoljar

Part VI Ethics and Politics 533

34 Lewis on Value and Valuing 535
Peter Railton

35 David Lewis's Social and Political Philosophy 549
Simon Keller

Bibliography of the Work of David Lewis 562
Index 572

Notes on Contributors

Helen Beebee is Samuel Hall Professor of Philosophy at the University of Manchester. Her research focuses on issues surrounding Humeanism and its rivals. She is the author of *Hume on Causation* (Routledge 2006) and *Free Will: An Introduction* (Palgrave 2013).

Karen Bennett is Associate Professor of Philosophy at Cornell University. She is the co-editor of *Oxford Studies in Metaphysics*, and the author of many articles in metaphysics and philosophy of mind. Her book *Making Things Up* is forthcoming with Oxford University Press.

David Braddon-Mitchell is Professor of Philosophy at the University of Sydney; he works in the philosophy of mind, metaphysics, and metaethics, and has published in these areas in various journals including *Mind*, *The Journal of Philosophy*, *Noûs*, *Philosophical Studies*, *Erkenntnis*, and *Synthese*.

Phillip Bricker is Professor of Philosophy at the University of Massachusetts-Amherst. He wrote his doctoral dissertation at Princeton University under the direction of David Lewis. He works primarily in metaphysics, especially issues in modality and ontology.

Rachael Briggs splits her time as a research fellow between the Australian National University and Griffith University. Her research interests include formal epistemology, metaphysics (particularly the metaphysics of chance), and preference-satisfaction theories of wellbeing.

John P. Burgess is the John N. Woodhull Professor of Philosophy at Princeton University, where he has taught since 1975. He is author or co-author of eight books and scores of papers and reviews in logic and related areas of philosophy.

John Collins completed a PhD at Princeton under David Lewis's supervision. He is Associate Professor of Philosophy at Columbia University. His current research centers on the nature of simple belief, the role of modal principles in epistemology, the foundations of causal decision theory, and the metaphysics of dispositions.

M. Eddon is an Assistant Professor of Philosophy at the University of Massachusetts-Amherst. Her primary area of research is metaphysics, with interests in fundamentality, quantity, mereology, and intrinsicality.

Alan Hájek is Professor of Philosophy at the Australian National University (since 2005). He works mainly in formal epistemology, the philosophical foundations of probability, decision theory, philosophy of science, metaphilosophy, philosophical logic, and philosophy of religion. He received his PhD at Princeton University, and worked for 12 years at Caltech.

Ned Hall teaches philosophy at Harvard University, and works primarily on topics in metaphysics and epistemology that overlap with philosophy of science (causation, laws of nature, objective chance and its relation to credence – all the fun stuff, in other words).

Katherine Hawley is Professor of Philosophy at the University of St Andrews, Scotland. She is the author of *How Things Persist* (Oxford University Press 2001) and of *Trust: A Very Short Introduction* (Oxford University Press 2012), as well as numerous papers within metaphysics and beyond.

Christopher Hitchcock is Professor of Philosophy at the California Institute of Technology. He has published extensively on the topic of causation, including articles in most of the leading philosophy journals, as well as venues in computer science, law, and psychology. He is also the co-editor of the *Oxford Handbook of Causation*.

Richard Holton is Professor of Philosophy at the University of Cambridge and a fellow of Peterhouse. He works in many different fields, and is the author of *Willing, Wanting, Waiting*.

Jenann Ismael is a Professor of Philosophy at the University of Arizona. She has published two books and numerous articles. Her research focuses on issues related to philosophy of physics including the nature of space and time, what quantum phenomena are telling us about the world, how fundamental ontology relates to higher level structures, and how we ourselves fit into the natural order.

Simon Keller is Professor of Philosophy at Victoria University of Wellington. He has published widely on topics in ethics and political philosophy. He is the author of *The Limits of Loyalty* and *Partiality*.

Rae Langton is Professor of Philosophy at the University of Cambridge. She works in ethics, metaphysics, feminist philosophy, and a range of other areas. She is author of *Kantian Humility: Our Ignorance of Things in Themselves* (Oxford University Press 1998) and *Sexual Solipsism: Philosophical Essays on Pornography and Objectification* (Oxford University Press 2009).

Ernie Lepore is Professor of Philosophy and Cognitive Science at Rutgers University. He has published in philosophy of language and philosophy of mind.

Stephanie R. Lewis taught philosophy from 1971 until 1984. When she realized that a tenured job was a complete impossibility, she went to Wharton and got an MBA. She has worked in public finance since then; nonetheless she is a philosopher first and last.

Fraser MacBride is Professor of Logic & Rhetoric at Glasgow University. He works on metaphysics and the philosophy of mathematics and is writing a book on the history of analytic philosophy. His recent publications include "How Involved Do You Want to Be in a Non-Symmetric Relationship?," *Australasian Journal of Philosophy*.

C.J.G. Meacham is an Associate Professor of Philosophy at the University of Massachusetts-Amherst. His main interests are in formal epistemology, decision theory, and the philosophy of physics.

Kristie Miller is an Associate Professor of Philosophy in the Department of Philosophy at the University of Sydney, Australia. She works primarily in metaphysics, in particular on the nature of time and persistence. Her most recent work focuses on the intersection of agency and timelessness.

Daniel Nolan is Professor of Philosophy at the Australian National University. He is the author of *Topics in the Philosophy of Possible Worlds* (Routledge) and *David Lewis* (Acumen/McGill-Queens), and articles in journals including *Noûs*, *Philosophy and Phenomenological Research*, *Philosophical Studies*, and *Analysis*. He works primarily in metaphysics.

Barbara H. Partee is Distinguished University Professor Emerita of Linguistics and Philosophy at the University of Massachusetts-Amherst. Her research centers on formal semantics; she is writing a book on the history of formal semantics. She also teaches semantics in Moscow and has worked with Russian colleagues on Slavic semantics.

Robert Pasnau is Professor of Philosophy at the University of Colorado. He works in the areas of metaphysics and knowledge, and especially the history of these subjects. He is the editor of *Oxford Studies in Medieval Philosophy*.

Peter Railton is Perrin Professor of Philosophy at the University of Michigan. David Lewis was his thesis supervisor. Railton's primary research has been in the philosophy of science, moral philosophy, and aesthetics. *Facts, Values, and Norms* (Cambridge University Press 2003) collects some of his papers in ethics and metaethics.

Craige Roberts is Professor of Linguistics and Adjunct Professor of Philosophy at the Ohio State University. Her work in formal semantics and pragmatics focuses on the nature of the context of utterance and the pragmatics of questions, presupposition, modals and attitude predicates, anaphora and reference, and their interactions in discourse.

Gideon Rosen is Stuart Professor of Philosophy and chair of the Council of the Humanities at Princeton University. He is the author (with John P. Burgess) of *A Subject with No Object* (Oxford 1997) and co-editor of the forthcoming *Norton Introduction to Philosophy*.

Jonathan Schaffer is Professor of Philosophy at Rutgers University. His research centers on metaphysics, epistemology, and language, and his publications include "Monism: The Priority of the Whole," "On What Grounds What," and "Knowing the Answer."

Wolfgang Schwarz is a postdoctoral research fellow at the Australian National University in Canberra. He works on topics in epistemology, metaphysics, philosophy of language, and logic.

Scott Soames is Distinguished Professor and Director of the School of Philosophy at USC. His recent books (from Princeton University Press) include *What Is Meaning?*, *Philosophy of Language*, *The Analytic Tradition in Philosophy Vol. 1*, and *Analytic Philosophy in America and Other Essays. Rethinking Language, Mind, and Meaning* is forthcoming.

Robert Stalnaker is Laurance S. Rockefeller Professor of Philosophy at MIT. He is the author of *Inquiry* (MIT Press 1984), *Our Knowledge of the Internal World* (Oxford University Press 2008), *Mere Possibilities* (Princeton University Press 2010), and two collections of papers, *Content and Context* (Oxford University Press 1999) and *Ways a World Might Be* (Oxford University Press 2003).

Daniel Stoljar is Professor of Philosophy at the Australian National University and an Australian Research Council Future Fellow. He is the author of *Ignorance and Imagination* (Oxford University Press 2006) and *Physicalism* (Routledge 2010) and co-editor of *There's Something about Mary* (MIT Press 2004) and *Introspection and Consciousness* (Oxford University Press 2012).

Matthew Stone completed his PhD in the Computer and Information Science Department at the University of Pennsylvania in 1998. Since then he has had an appointment in the Computer Science Department and Center for Cognitive Science at Rutgers, the State University of New Jersey. Stone has had visiting positions at the University of Edinburgh and the Universität Potsdam. He works on problems of meaning in human–human and human–computer conversation.

Brian Weatherson is Marshall M. Weinberg Professor of Philosophy at the University of Michigan. His recent research is on the philosophical significance of normative uncertainty, and on the role practical and theoretical interests play in the connections between knowledge, belief, and credences.

J.R.G. Williams is Professor of Theoretical Philosophy at the University of Leeds. Working in the philosophy of logic, language, metaphysics, and formal epistemology, he directs the ERC Nature of Representation project. Publications include "Decision Making under Indeterminacy," *Philosophers' Imprint* (2014), "Counterfactual Triviality," *Philosophy and Phenomenological Research* (2012), "Eligibility and Inscrutability," *Philosophical Review* (2007).

Jessica Wilson is Associate Professor of Philosophy at the University of Toronto. Her primary research interests are in general metaphysics (especially modality and indeterminacy) and the metaphysics of science (especially inter-theoretic relations). Recent publications include "What Is Hume's Dictum, and Why Believe It?," *Philosophy and Phenomenological Research* (2010), "Fundamental Determinables," *Philosopher's Imprint* (2011), and "A Determination-Based Account of Metaphysical Indeterminacy," *Inquiry* (2013). She is writing a book titled *Metaphysical Emergence*.

Part I

Biography and New Work

1

Intellectual Biography of David Lewis (1941–2001)

Early Influences

STEPHANIE R. LEWIS

This chapter is not a cradle-to-grave intellectual biography of David Lewis. In particular, it does not try to be comprehensive about the origins of his views or of how he came to hold them. Its purpose is to exhibit elements of the origins of the David Lewis we knew, philosopher and human being, and whose works we know. It describes important influences on David as a child, as an adolescent, and as a young man.

Let me begin with the last, and most important, of the forces that shaped the adult David, and made him the philosopher that he was. Not the only influence: nothing would have made David into the philosopher he was if he didn't have the wherewithal to begin with.

David, and usually I as well, made many visits to Australia: in 1971, in 1976, and nearly every year (except 2000, the year of David's kidney transplant) from 1979 right through 2001. He gave talks, went to talks, conversed with many people, and whenever he was in the right place at the right time he attended the Australasian Association of Philosophy conference. We toured around and enjoyed the urban amenities of Melbourne and, to a lesser extent, Sydney. And, starting in 1980, we went to the footy (Figure 1.1). David, who in general had no interest whatever in sport, somehow became a one-eyed supporter of the Essendon Football Club, in the Victorian (subsequently, the Australian) Football League. He was buried with his Essendon membership card in his pocket.

In July of 2002, nearly a year after David's death, I visited Australia by myself, and attended the Australasian Association of Philosophy conference. At the conference dinner, someone rose and asked us all to take a moment to remember David. After a minute or so, they shoved the microphone at me and asked me to say something. My only preparation for this was three glasses of wine. The first words that came out of my mouth were "Australia made David." I must have said more, but I have no recollection at all of what it might have been.

1.1 Childhood

David was born into an academic household in Oberlin, Ohio, on September 28, 1941. He was the eldest of three children. His father, John D. Lewis, was professor of government at Oberlin College,

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

Figure 1.1 Going home from the MCG after a loss to Collingwood, 1984. © Stephanie Lewis 2013.

where he taught from 1936 until his retirement in 1972. John was one of the great Oberlin teachers of his time, an Oberlin oligarch. As student and faculty member, he was at Oberlin for 41 years. John hadn't much standing as a scholar or researcher, especially in the later part of his career; his mark was on his generations of students, including Cecilia Kenyon, Kenneth N. Waltz, Sheldon Wolin, and W. Carey McWilliams Jr. Many others in other careers expressed their gratitude for his intellectual influence on them.

David's mother, Ewart Kellogg Lewis, was the scholar, by inclination, anyway. Unlike her husband, she came from an academic family. She was a graduate of the University of Wisconsin, also Phi Beta Kappa, having fled Wellesley College, and she held the PhD from the University of Wisconsin. She published *Medieval Political Ideas* (1954), a collection of critical translations and introductory essays of medieval philosophers and political theorists. She published no other scholarly works that I can find. Her reputation in medieval political theory survives.

She had no formal teaching career to speak of. Oberlin had, or was thought to have, a nepotism rule, so, other than casual employment in the history department, a post at Oberlin College was denied her. Neither she nor her husband was inclined to challenge this and, in any case, her employment at Oberlin ended after a squabble over the appointment of another faculty spouse. She did have an instructorship at Western Reserve University in Cleveland for three years.

She was an academic to the core, even though running the household fell to her. She taught her children to read early, and strongly encouraged David's native bookishness. When David was nine or ten years old he had an attack of polio, and, unrelated to this, a bone cyst in his thigh was discovered. He had a transplant of bone chips to cure the cyst, and as a consequence spent several weeks in, or mostly in, bed. Ewart taught him Latin. (He also took Latin in high school.) She also taught him to type properly.

David, born of two Phi Beta Kappa academics, was the eldest of three siblings. Being the eldest, and a little ungovernable, and being recognized from an early age as someone with intellectual curiosity and motivation, he was allowed to follow his own inclinations about his studies and activities.

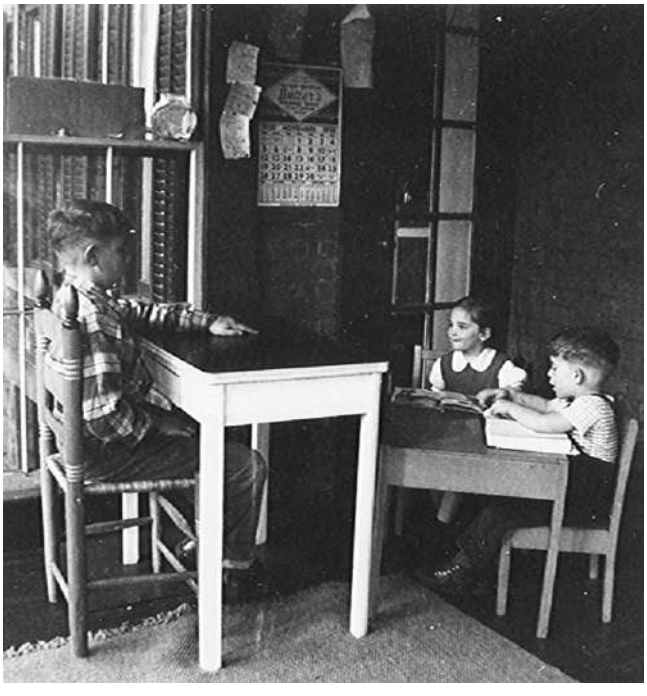

Figure 1.2 Photo by Ewart Lewis, November 1950.

The portion of this chapter dealing with David's childhood and early adolescence draws partly on Lewis family myth and folklore, but primarily on an autobiography he wrote, at the age of 14, in his next-to-last year of high school. It doesn't show much introspection: it has a lot of facts and family history in it. But it does describe his interests at various times. He, like most smart kids, read a lot and was interested in science. The autobiography has next to nothing in it about school friends, and most of the stories of family interactions are about his father. He was a solitary boy, planning and doing projects by himself, and reading. From what he says about various science projects his attention span appeared, even as a small child, to be unbounded. He wasn't unsocial but, if the autobiography is accurate, most of his interactions were with adults. There is only one mention of a friend of his own age.

It isn't as if David didn't care for his siblings, nor they for him. There is a photo of David from 1950, when he was nine years old, sitting in his father's study at their house, teaching school to his brother Donald, then five years old, and his sister Ellen, then three. They are listening raptly, their books open before them. The posture of David explaining something to an attentive audience will strike anybody who knew him as familiar (Figure 1.2).

For all practical purposes, David barely went to high school. Between the fall of 1954 and June of 1957, his high school years, he attended several courses at Oberlin, General Chemistry and Organic Chemistry among them, and took the exams and did the lab work. He had a chemistry lab in the basement of the Lewis house, where he did chemistry experiments and glassblowing. (And no, he never did nearly blow up the house.) One summer he worked on a project in a college lab, supervised by Professor Renfrow, in the Oberlin chemistry department. David Sanford remembers him from Oberlin chemistry classes as smarter and better prepared than any of the other students.[1]

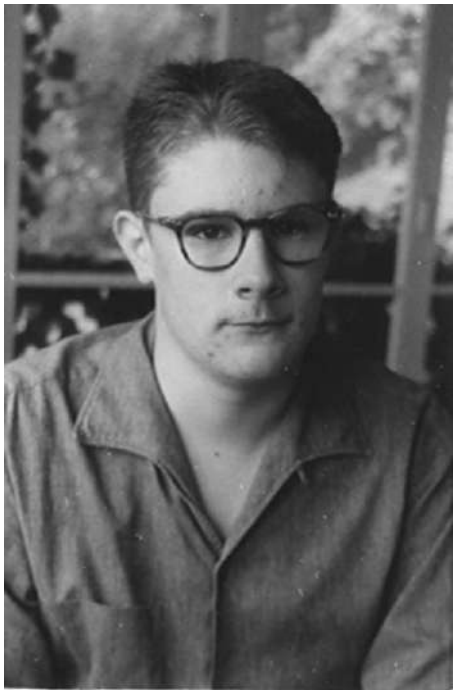

Figure 1.3 David Lewis at age 14. Photo by Ewart Lewis.

David also showed early signs of the highmindedness that characterized him for his entire life. In a draft of his essay to accompany his application to Swarthmore College, written in the spring of 1957, he says:[2]

> Last spring [of 1956], when a high school teacher was fired without reasons given, I was one of five students who drew up and circulated a petition asking the [Oberlin] Board of Education to give reasons. This petition, signed by about 60% of the High School students, was followed by a series of petitions and protests by teachers and citizens which finally resulted in a thorough investigation of the school situation by the Board of Education, and the replacement of the Superintendent of Schools by a new man who is initiating several much-needed reforms, I got very much interested in the whole situation and have been attending School board meetings regularly since then.

In the course of high school, his interests evolved and he continued to grow into the David we knew. Here is the last section of that autobiography: he was 14 when he wrote it, in 1955 [Figure 1.3], and there is no evidence whatever of ghostwriting (Ewart did type it) by either his mother or his father. David is uncharacteristically pompous, but the voice is his own.

> LOOKING AHEAD
>
> This, then, is the story up to now. But it is still incomplete. After all, one's first fourteen years are not the greater part of life; it is necessary to say something about the future. Moreover, this has been a record

primarily of events: I have not yet said much about what I think of it all. And these are important; for an event is almost meaningless, as I see it, in comparison with an idea.

To take first the matter of concrete plans for the future, I must begin by saying that I am not sure of any of it. I expect to finish high school in the next year, taking, perhaps, some courses at Oberlin College also. After that, I intend to enter some college; I do not know where. I would like to go to a serious college, preferably a small one, where I can devote my time to work without being made to seem abnormal by doing so. I hope to get a complete liberal education, not just a technical one. It is for this reason that I am doing so much college work now; the programs for one majoring in a technical field in college are all too often so time-consuming that I would not be able to work on anything else. I expect to concentrate on technical studies in graduate school, but in college I want some freedom in arranging my schedule. I have been starting recently to consider the choice of a college, but this has been very difficult. The only place that seems to meet my requirements is Oberlin, and my parents and I agree that I should go away from home for college.

I am not sure what I want to study. Until very recently I expected to concentrate in chemistry, but I am losing some of my interest in it now. I do not know whether I am actually becoming tired of it simply because I have been concentrating on it so heavily for the last three years, or whether it is really not my proper field. Of course, it may also be that I am simply reacting to the very dull lectures which I hear at my chemistry class at the college.

I have many other things besides chemistry which I am interested in. I particularly like mathematics. Its logic, in particular, appeals to me. I do not like the mechanical processes where one puts in numbers and "turns the crank" to find a solution, nor do I like the problems of applied mathematics. But the basic concepts, the logic, the reaching of conclusions from reasoning alone, these are for me. I am also interested in some of the ideas of philosophy, metaphysics especially, although I have never yet made a detailed study of them. I have done some thinking of my own along these lines, and, of course, my results turn out to have been around for centuries.

I am interested, though not quite so much, in several other subjects. For instance, during my freshman year in high school I became very much interested in Latin and in the history of the Romans. I also am interested in other sciences, with the exception of biology, which leaves me quite bored. Probably the reason for this is that biology seems to me to be just a jumble of dull information, whereas the other sciences are logical structures.

After college I do not know just what I want to do. If I specialize in science, I will be able to get all sorts of industrial jobs. But this does not appeal to me. It is all a matter of intensely practical, routine work. My interest is not here so much as it is in the theory. Perhaps the best thing would be to enter the field of college teaching. Here I could work on any project I pleased. I would also be with people of my own sort. If I decide to work in some other field, I can say nothing of what I would be doing.

And now for the ideas I have been able to gather. My philosophy is, more than anything else, philosophy in the literal sense, love of knowledge; but not just knowledge; of understanding the realities of the universe, the reasons for everything. For I feel that there must be a cause for everything, that the past determines the future, that there must be certain natural laws, or perhaps only one, such that it would be possible to deduce all the features of the universe from it alone. This is a scientist's way of looking at the universe; I do not deny that there may be other ways which can reveal truths unknowable from the viewpoint of pure reason. Religion is such a way, so perhaps is art. My feeling about religion is that I cannot accept the elaborate system of details which an organized religion tends to build up, that it is in conflict with all reason. I do not find such a simple solution, though, to be the question of the existence of some kind of a God. My attitude is that it is impossible to prove or disprove the existence of God by logic. Indeed, not only impossible, but very foolish, as though I were to try to prove mathematically that my eyes are blue [they were]. The existence of God and the domain of logic are, I feel, absolutely separate. These are my views; I would not be so dogmatic as to say that there is any reason whatever to consider them correct.

I can see no other meaning in life except the gaining of knowledge; but this is rather meaningless, a rationalization of the fact that I have devoted myself so much to the gaining of knowledge, to the over-exclusion of other aspects of life. I can see that this is not desirable and I am trying to do something

about it, although it is a slow and rather difficult business. Nevertheless, I feel that I am making some progress.

I feel that the world is good, although it is sometimes hard to see. But on the whole, I think it is good. As for my feelings on the problems of the world, I am rather idealistic. I think this is right, but it puts me somewhat at a disadvantage in practical affairs.

* * *

These are the events and the ideas of my life until now; it remains for the reader to judge them.

The End

David finished his senior year of high school but did not get a diploma: a civics requirement or something had not been met. So he was not a high school graduate, and thus was barred from serving in the Ohio National Guard (not that he wanted to). He did get a Merit Scholarship. Shortly before he turned 16, in September of 1957, he began his freshman year at Swarthmore College.

1.2 Swarthmore: The First Two Years

David had been solitary, though apparently not lonely, as a child. At Swarthmore he made friends, pretty much instantaneously. Several of these friendships endured for the rest of his life. Among his contemporaries were many future philosophers: among them, Allan Gibbard, Gil Harman and Peter Unger, and the linguist Barbara Hall, later Partee. He quickly became a part of the Swarthmore folkie scene.

He started out with the intention of majoring in chemistry, but took philosophy classes as well: in his sophomore year, he took Intro to Philosophy and Symbolic Logic. Most of his grades were As, but he did get some Bs in science courses.

Between his sophomore and junior years, the year he turned 18, the family spent a year at Oxford University. This was the year 1959–60. John had a research fellowship; though no publications followed upon it. He drove his new Jaguar sedan every day from their house in Wheatley to St Catherine's Society, which was not yet a college. He wore tweeds and a "Toad of Toad Hall" cloth cap. Ewart didn't have any college appointment. She read and conversed widely, and had a big part in the family excursions to various places in England and France, but her role was at home.

1.3 Oxford

David took a break from chemistry and physics. He did philosophy. He was treated by his father's college, St Catherine's Society (as it then was), as an undergraduate student, and assigned to the college philosophy tutor, John Simopoulos, who took one look at him and handed him off to Iris Murdoch. She was then a philosophy tutor at St Anne's College. He wrote weekly papers for her and discussed them in the one-on-one tutorials which were than a part of an Oxford undergraduate's study. He also attended lectures by, among others, Grice, Strawson, Ryle, and J.L. Austin. Despite repeated exposure to the ordinary-language culture of Oxford at the time, David never caught the disease.

Murdoch wrote letters of recommendation for David. Here is what she said about him (she sent him this letter also):

ST ANNE'S COLLEGE
OXFORD
Telephone 57417

June 15 1960

Mr David Lewis has worked with me on moral philosophy for the best part of three terms during his stay in Oxford. About half of this time was spent on studying traditional philosophers (Hume, Kant, also Mill, Moore and others) and the other half on looking at contemporary theories, especially in relation to freedom, and in discussing versions of Mr Lewis' own ideas on the latter subject. There is no doubt that Mr Lewis is a very gifted young man indeed and has a true talent for philosophy. Were he to remain in that subject (which unfortunately appears to be unlikely) I should advise him to forget about his own theories for a while and spend time grappling with difficult and unfamiliar ideas in the great philosophers of the past. This however is to say no more than that Mr Lewis is young and (naturally) still in need of education. His own ideas in fact are both interesting and original. His work has been excellent, certainly "alpha" throughout.

Iris Murdoch
Fellow and Tutor

(Reprinted by kind permission of Iris Murdoch.)

And here is a remark she made in a letter to him, about graduate school letters of reference she had been doing for him (reprinted by kind permission of the Estate of Iris Murdoch):

St Anne's College
Oxford
Nov 23 1961

Dear David,
Thanks for your letter [there is no copy of this letter in David's files]. I've sent off the two forms. (How is your emotional maturity nowadays? I found it hard to think of a mark for that.) I will do the other two testimonials very soon. I think I could commend Black & Malcolm for my (indirect) knowledge of them. Mrs Foot lately has had a very lively & profitable sojourn at Cornell. Anyway, best of wishes. I see you don't list Yale, rightly, I'm sure. Bulldog, bulldog etc.

Yours,
Iris

(Reprinted by kind permission of Iris Murdoch.)

1.4 Swarthmore: Second Two Years

David returned to Swarthmore as a philosophy major. (Swarthmore wouldn't let him do philosophy as a related minor along with a science major.) He did take advanced calculus, and two more physics courses, and he attended a course in linguistics at Penn with Henry Hiz. But it was otherwise all philosophy. He either took or sat in on most of the courses offered by the small philosophy department at Swarthmore. Jerry Shaffer was among David's favorite teachers. David continued to work in moral philosophy: a research project, supervised by Monroe Beardsley, resulted in *Can Ethics be Reasonable?*, a 40-page essay, typed, single-spaced, with practically no margins. (David kept a copy, which I have.) The typewriter is the one he used for the rest of his life.

In the essay he says that "personality is the sum of actual and potential behavior…" most of which patterns of behavior are "ideal forms of life." Ethical thought consists in "compositing,

presenting, elaborating, and relating these ideal forms of life."[3] To this view he adds: "To this model must be added social morality, a law-like system of regulating overt acts so as to reconcile conflicting interests of different people."[4] He concludes the essay thus:

> To summarise: "good," "right," "ought," "should," "duty," "obligation," "fortunate," "harm," "benefit," "desirable," "wicked," "immoral," "choiceworthy," (if there is any such word) etc. get their meaning among some or all of the following components:
>
> 1. Expression of decision; "internal motivation."
> 2. A small range of purely descriptive meanings, logically independent of attitude or conduct.
> 3. In the case of "good" and perhaps "right" a large range of attributive descriptive meaning, at least sometimes with dependence on attitude and conduct on pain of irony or inappropriateness, if not of falsehood.
> 4. To go proxy for reasons within a context of purpose. Perhaps with the implication on pain of irony or inappropriateness that the purpose is itself supported by empathically understood ideals.
> 5. To go proxy for reasons, empathically understood ideals, which for some reason there is no need to detail. Some of these words are limited to going proxy for only certain classes of ideals. If these words have a purely ethical sense, this is it.[5]

He acknowledges the strong influence of several of Murdoch's published works in philosophy on the essay.

He graduated from Swarthmore in 1962, with high honors in philosophy and election to Phi Beta Kappa.

1.5 The Hudson Institute

Between 1962 and 1975, David was a part-time member of staff at the Hudson Institute, a policy think tank then concerned with issues in strategy and deterrence and nuclear weapons policy. He worked on research assignments, often with Herman Kahn, then Hudson's director, and with Max Singer. Hudson now has the reputation of being a conservative organization; then it was more concerned with the technical aspects of nuclear weapons, disarmament policy issues, and Kahn's interest in what Kahn called "futurology." David developed an interest in game theory as a result of the Hudson work.

1.6 Graduate School: Australia I, Jack Smart

David arrived at Harvard in 1962, bringing with him his native mind–body identity theory, space–time four-dimensionalism, his Humeanism about causation, his Rylean behaviorism, his immunity to ordinary-language philosophy, and his disdain for political correctness. In the fall of 1963 he attended a seminar given by J.J.C. Smart, then visiting Harvard from Adelaide, on "Philosophical Problems of Space and Time." This was the beginning of a friendship that lasted nearly 40 years, until David's death in 2001 (Figure 1.4). They conversed about philosophy and exchanged letters on philosophy.

Jack invited David to give the Gavin David Young lectures at Adelaide University in August of 1971.

At Harvard, David worked with W.V. Quine and with Hilary Putnam, going to MIT for Putnam's seminars before Putnam moved to Harvard. He also went to linguistics seminars at MIT. He admired

Figure 1.4 David Lewis and Jack Smart. Belcunda, South Australia, August 1971. © Stephanie Lewis.

Quine's philosophical style and prose style as well; David learned to write, and speak, clearly and concisely about even the most technical matters in philosophy. He used to quote C.G. Hempel with approval: "it adds neither to the rigor of the argument nor the clarity of the exposition to say that a man *M* crossed the street *S*."

David's thesis, *Conventions of Language*, grew out of his interest in language, encouraged by Quine, and also out of David's own continuing interest in game theory. Metaphysics was not yet at the center of his interests. The game theory aspects of his theory of convention owed a lot to his Hudson Institute work and to conversations with Thomas Schelling. Schelling's prose style, in *The Strategy of Conflict* (1960), also served as a model for David's writing.

1.7 UCLA

David and I went to UCLA in 1966, he as an assistant professor and me as a graduate student in philosophy. He got interested in formal semantics, and had many discussions with Richard Montague and with Barbara Hall Partee. David went to Montague's seminars on formal semantics. He and Hans Kamp had a lot of discussion on the subject, and to a lesser extent David conversed with David Kaplan as well. And all the while he wrote and he wrote and he wrote.

1.8 Australia II: David Armstrong

In 1968 David Armstrong was visiting at Stanford. He and David Lewis met and talked. This began a friendship and a philosophical interchange, carried on in letters and conversations, for the rest of David's life.

1.9 Australia III: The First Visit

When we arrived in Sydney in July of 1971, we stayed for several days with David Armstrong and his first wife, Madeleine. This was the beginning of the philosophical conversations between the two Davids.

In August of 1971 David delivered the Gavin David Young lectures at Adelaide University. The topic was time travel. This was the first of, if I have counted right, 26 visits. This was for two months, and gave David the chance to go around to most of the philosophy departments at Australian universities – we missed Perth – and give talks, go to talks, and get to know people. And he wrote and he wrote and he wrote.

The effect on David cannot be overstated. He found many friends, and many, many opportunities to discuss matters of common interest. And he enjoyed himself, not least because he didn't have any teaching or departmental responsibilities. Like the rest of his Australia trips, it amounted to study leave.

1.10 Princeton

David arrived at Princeton in the fall of 1971, where he taught for the rest of his life. This essay looks at David's early life, and thus doesn't say much of anything about his Princeton career. It is worth mentioning that he had over 30 PhD students, either entirely or partly under his supervision, many of whom now ornament the profession.

1.11 Australia IV: 1976

David made his second visit to Australia in July and August of 1976. David and Madeleine Armstrong invited us to spend a weekend with them at Glenogil, the house of their friends Pat and Rosemary Ryan, near Avenel, Victoria. The two Davids walked, and talked, for a large part of the daylight hours of two days. This visit firmly established the friendship between the two Davids (Figure 1.5). They never agreed about much: DKL was by then a modal realist, DMA had no use at all for more than one possible world. They disagreed fundamentally about universals, and about properties. But they never stopped talking about philosophy. The Lewis–Armstrong correspondence is by far the longest, and richest, and most detailed of all of David Lewis's correspondences.

As a result of the Australia experience, which continued until 2001, three months before he died, David became a lot less solitary, both socially and intellectually. He learned how to have (or, better, became accustomed to having) sustained philosophical discussions: on walks, in seminars, at people's houses, at parties, in the department tea room, and in the pub. This didn't result in a lot of agreement, but it gave all parties a chance to articulate and refine their positions.

And he wrote letters, from his early college time until three days before he died. He never used email, so there are no ephemera. He kept the letters he received, and he nearly always kept copies of what he wrote. The archive of his correspondence includes many thousands of philosophical letters, and many threads of philosophical discussion. These letters, many of them, also portray friendships.

David was, from his earliest childhood, capable of sustained attention to what interested him, and as he grew up his writing style developed into the lucid and witty prose we know. What Australia, and the Australians, did for him was to take him out of himself and make him into a member of a community (Figure 1.6).

Figure 1.5 The Two Davids, Glenogil Station, Victoria. August 1976. © Stephanie Lewis.

Figure 1.6 North Queensland, 1990. © Stephanie Lewis.

Acknowledgments

Thanks are due to many people for reminiscences, discussion and comments, especially to John Cooper, Allan Gibbard, Wendy Goble, Alan Hájek, David's brother Don Lewis, his sister Ellen Lewis, Barbara Hall Partee, David Sanford, and to Jonathan Schaffer. David's Swarthmore College transcript, along with copies of parts of his college applications, exist, and provided me with insight into David's character and personality at a time before I knew him.

Notes

1 Pers. comm. 2012.
2 This draft, along with his Swarthmore College transcript, was filed away in his "Swarthmore" folder. David never threw away a piece of paper in his life if he could help it. Well, hardly ever.
3 Summary of § 2.1, p.1.
4 Summary of §3.1, p.1.
5 Conclusion of the essay, §3.5, p. 38.

References

Lewis, E.K. (1954). *Medieval political ideas*, 2 vols. London: Routledge and Kegan Paul, reissued 2012.
Schelling, T. (1960). *The strategy of conflict*. Cambridge, MA: Harvard University Press.

2

Counterparts of States of Affairs[1]

DAVID LEWIS

Counterpart theory affords an especially flexible form of essentialism. By considering the very same thing under different, equally legitimate counterpart relations, we can endow it with different essences and different potentialities. By endowing ordinary things with peculiar essences, we could provide truthmakers for intrinsic predications (Lewis, 2003). By endowing the entire world with peculiar essences, we could provide truthmakers for negative existential propositions (Rosen and Lewis, 2003). Thus we avoid the need for states of affairs or nontransferable tropes as truthmakers. And a good thing too, I used to think, because these alleged entities are involved in *prima facie* mysterious necessary connections between distinct existences.

But the same technique for endowing things with tailor-made essences by suitable choice of a counterpart relation can be applied also to states of affairs themselves or to tropes. When we do, we render the necessary connections unmysterious. So I've come around to thinking that a theory of states of affairs or tropes, assisted by flexible counterpart theory, is after all another entirely satisfactory way to provide truthmakers.

States of affairs are somehow constructed from particular things and the properties they instantiate. A familiar argument says that this construction must be neither mereological nor set-theoretical: else the state of affairs will exist if the thing and the property do, never mind whether the thing instantiates the property, so the state of affairs is not a truthmaker for the proposition that the thing instantiates the property. If so, the need for some third novel sort of construction is a further drawback of a theory of states of affairs. However, that familiar argument rests on a thesis of mereological or set-theoretical essentialism. If essences can be tailor-made to suit our purposes, we need not be saddled with *any* unwelcome sort of essentialism.

The thesis of mereological essentialism says that it is essential to something that it has exactly the parts that it actually has; and conversely that it is essential to the parts that they compose exactly the mereological sum that they actually do. So if a train, say the 15:40 from Paddington yesterday, consists of two carriages, DMBS1234 and DTS6789, then we have a puzzle. The sum DMBS1234 + DTS6789 seems to be necessarily composed of DMBS1234 and DTS6789; yet the train

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

could very easily have been composed of different carriages, since carriages are interchangeable and are routinely taken out of service for repairs; and yet the train *is* the sum! Counterpart theory to the rescue. There's one legitimate counterpart relation that validates mereological essentialism: it obeys the rule that the counterpart of a sum is the sum of the counterparts of the parts; so for instance the counterpart of DMBS1234 + DTS6789 is the sum of the counterparts of the carriages DMBS1234 and DTS6789, never mind whether those counterpart carriages are coupled together into a train. And there's another equally legitimate counterpart relation on which the counterpart of yesterday's 15:40 from Paddington is yesterday's 15:40 from (the counterpart of) Paddington, never mind what carriages it consists of. The train *is* the sum. Yet if we call it a sum, that tends to evoke the first of these counterpart relations, while if we call it a train, that tends to evoke the second. We could put it this way: this one thing qua sum has one essence, the same thing qua train has another. Qua sum it essentially consists of just these carriages, qua train not.

Much the same can be said of a mereologically constructed state of affairs: a thing-plus-property sum such that the thing instantiates the property. It is the state of affairs *Fa* of *a*'s having *F*; it is the sum $a + F$. Calling this one thing a sum evokes one counterpart relation, one that validates mereological essentialism. Calling the same thing a state of affairs evokes another counterpart relation, one that doesn't validate mereological essentialism, but does obey the rule that any counterpart of a state of affairs must be a state of affairs. (Unlike the train-counterpart relation, it does validate one direction of mereological essentialism. If *Fa* is mereologically composed of *F* and *a*, any state-of-affairs counterpart of *Fa* must be mereologically composed of *F* – or perhaps a counterpart of *F* – and a counterpart of *a*.) Here in this world we have one thing which is *Fa* and $a + F$. Off in some other world we have a counterpart a' of *a*, and we again have property *F* (or maybe we have a counterpart of *F*); but a' doesn't instantiate *F*, so the sum $a' + F$ is not a state of affairs. Our thisworldly state of affairs/sum has $a' + F$ for a sum-counterpart, but not for a state-of-affairs-counterpart. Qua sum, it is essentially the sum of *a* and *F*, but not essentially a state of affairs, and therefore unfit to serve as a truthmaker for the proposition that *a* has *F*. Qua state of affairs, this same entity *is* essentially a state of affairs, and therefore a truthmaker for the proposition that *a* has *F*.

If we prefer to construct states of affairs set-theoretically, say as thing–property ordered pairs such that the thing instantiates the property, exactly the same treatment applies. We have one legitimate counterpart relation that validates both directions of set-theoretical essentialism: any counterpart of the pair of *a* and *F* is a pair of a counterpart of *a* and a counterpart of *F*, and conversely, never mind whether the thing instantiates the property. The pair qua pair is essentially the pair of *a* and *F*, but not essentially a state of affairs, and therefore unfit to serve as a truthmaker. We have another legitimate counterpart relation which validates only one direction of set-theoretical essentialism, but under which any counterpart of a state of affairs must be a state of affairs. The state-of-affairs-counterpart of *Fa* must be a pair-counterpart of the pair of *a* and *F*, but a pair-counterpart of the pair is a state-of-affairs-counterpart of *Fa* only if it is a state of affairs – that is, only if its first term instantiates its second. The state of affairs qua state of affairs is essentially a state of affairs, and so is a truthmaker for the proposition that *a* has *F*. The very same thing qua pair is not.

Suppose there really is some other kind of construction, neither mereological nor set-theoretical – symbolize it by *. We could say that a state of affairs *Fa* is a *-structure ($a * F$) such that *a* instantiates *F*. It's plausible that, despite our new kind of construction, the situation is the same again. Qua *-structure, ($a * F$) is essentially *-constructed from *a* and *F*, but not essentially a state of affairs (unless for some reason *-construction only works when the first term instantiates the second). Qua state of affairs, *Fa* is essentially a state of affairs, and (if it has essentially the *-constituents *a* and *F*) it is a suitable truthmaker for the proposition that *a* has *F*. And yet ($a * F$) is the very same thing as *Fa*.

Another sort of entity said to be associated with predications is a trope. This time, construction goes in the other direction: ordinary things and properties are constructed out of tropes, not *vice versa*. Nevertheless, it is well known that if we want tropes to serve as truthmakers for predications, they have to be *non-transferable* tropes. Let thing *a* have property *F*; let *t* be *a*'s *F*-ness trope. If *t* could exist apart from *F*, *t* might exist even if *a* lacked *F*, or even if *a* did not exist at all. (Non-transferability might mean that *t* is inseparable from the rest of the trope-bundle that constitutes *a*, or it might mean that *t* is inseparable from the particular substrate of that bundle. Let us leave that question open.) And non-separability is another *prima facie* mysterious necessary connection.

Counterpart theory to the rescue again. If essences in general can be tailor-made by suitable choice of a counterpart relation, essences of tropes can be tailor-made. For purposes of truthmaking, we would of course want an *F*-ness trope to be essentially an *F*-ness trope; it had better not have counterparts that are *G*-ness tropes instead. Likewise, we would want *a*'s *F*-ness trope to be essentially *a*'s; its counterparts should belong to counterparts of *a*. Perhaps for other purposes we'd want a counterpart relation that worked differently – as it might be, one that gives more weight to the place of a trope in the wider world. But as usual we can have multiple counterpart relations, evoked by different ways of referring to the same entity and by other features of conversational context.

Conclusion. If we help ourselves to the flexibility of counterpart theory, ordinary things can serve as truthmakers. States of affairs or tropes are not needed for purposes of truthmaking (whatever other purposes they may serve). But also, if we help ourselves to the flexibility of counterpart theory, the principal difficulties that stand in the way of using mereologically or set-theoretically constructed states of affairs, or tropes, as truthmakers vanish. The requisite necessary connections lose their mystery; and we no longer need an unfamiliar new kind of construction that is neither mereological nor set-theoretical.

Acknowledgments

I thank Phillip Bricker and Cian Dorr for suggestions that led to this chapter.

Note

1 Draft, December 14, 2000, transcribed by Anthony Fisher, September 2014.

References

Lewis, D. (2003). Things qua truthmakers. In H. Lillehammer & G. Rodriguez-Pereyra (Eds.), *Real Metaphysics: Essays in Honour of D. H. Mellor*, pp. 25–38. London: Routledge.

Rosen, G. & Lewis, D. (2003). Postscript to 'Things qua truthmakers': negative existentials. In H. Lillehammer & G. Rodriguez-Pereyra (Eds.), *Real Metaphysics: Essays in Honour of D. H. Mellor*, pp. 39–42. London: Routledge.

3

Reply to Dana Scott, "Is There Life on Possible Worlds?"[1]

DAVID LEWIS

The triumphal advance of the beings from logical space seems less formidable from the inside. Some of those who served the cause best a few short years ago have turned away. Others fight on, but take care to keep open a line of retreat. Now and then they rest from their labors to say that possible worlds are, after all, only a metaphor or a useful fiction. A few (Carnap, Adams) then say what they take to be the literal non-fiction – fair enough! But others keep their real views very much to themselves. Others stick to the pure and safe model theory of possible- worlds semantics, never quite speaking of possible worlds but only of an index set *I* which may be regarded as the set of possible worlds. And then there's the second front: how shall we deal with those enthusiasts who are even now preparing to bring on the *im*possible worlds? But enough grumbling – to work.

What are the alternatives to realism about possibles? How about taking modality as primitive? I said that would be no theory at all, but abstinence from theorizing. I've been quite properly scolded, on this and other occasions, for seeming to suggest that there's no way ever to theorize about something without reducing it to something else. That's false: set theory, for instance, is an illuminating but non-reductive way of theorizing about sets. Still, I don't think I was so wrong about this case. Taking modality as primitive has led, in practice, to trial-and-error modal logic. Theses do not a theory make, not even when they're neatly organized in a formal deductive system; explanatory power also is needed. I have no analysis of explanatory power to offer, but I think that I know it when I see it, and I miss it when I don't. Consider the problem of explaining why some superficially plausible modal inferences are fallacious. In trial-and-error modal logic, the game is to issue provisional decrees: thou shalt not infer according to rule XYZ in modal contexts! You win if you stop all the paradoxes so far discovered. But don't ask why you should give up rule XYZ, which still seems as if it ought to be truth-preserving. The only answer is that it seems somehow to get you into trouble. That's what I call abstinence from theorizing. I much prefer the way possible-world semantics has handled the same problem. Hypotheses are put forward, under which the modal fallacies turn out to be instances of more general fallacies that we already understood. We know just why rule XYZ let us down; and, as a bonus, we know just what repaired version would escape the trouble.

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

Pure model-theory, I think, cannot perform this service. It can very well show us how an inference fails in a certain class of interpretations. But unless it is at least hinted that the actual interpretation falls within that class, I don't see how anything whatever has been done to explain the actual failure.

Fortunately, trial-and-error modal logic is not the only way to take modality as primitive. There is also the option of reducing possible worlds to maximal consistent "books" of propositions. As a reductive proposal this is circular, because "consistent" had better mean "compossible." As a non-reductive way of theorizing about modality, however, it may have merit. The books may be enough like possible worlds so that at least some of the explanatory power of possible-worlds semantics carries over – my guess would be that there is explanatory power for matters of *de dicto* modality, but not for those areas of *de re* modality that my favorite form of possible-worlds semantics either explains or mis-explains, as the case may be. But is this reductive option credible? I find it absolutely incredible that our actual world is a maximal consistent book of propositions! When I wrote *Counterfactuals* I thought that was enough to settle the matter; but Hartry Field and Robert Adams and someone I believe to have been David Kaplan have led me to think again. What you must do to maintain the reduction of worlds to books is, I think, to declare that the actual world is not a possible world. There are the possible worlds, and they are books of propositions, and one of them is the actual possible world, and it is the book made entirely of true propositions; but the actual world itself is something else again. I have argued elsewhere that the actual world should not be regarded as different in kind from the otherworlds, on pain of making it mysterious how we know that we are actual or (alternatively, according to a suggestion of Robert Adams) on pain of denying that it is a contingent matter which world is actual. The reply could be that that's right as applied to the actual possible world, i.e. the actual book; but wrong as applied to the actual world itself. I see no reason to prefer this position to my own, but also no knock-down argument against it.

I move on, now, to Professor Scott's misgivings on another score: *license to metaphysical creation*. I really must insist, at the outset, that to hypothesize that there are *X*s is not to create them! I don't suppose Scott would ever confuse hypothesizing that there are Reds under beds, or perfect squares with an odd number of factors, or quanta of gravitational interaction, or God, with creating any of these things! Yet the dangerous metaphor of hypothesizing as creating pervades his paper, and sometimes I find it hard to see the literal meaning behind the metaphor. For instance when he says "It looks to me that the totality of propositions is only a very potential one" I'm pretty sure he doesn't mean that we can go on creating more and more propositions forever with nothing to stop us; but what does he mean? And what is he getting at when he says that my reliance on belief (i.e. my appeal to naive opinion as a reason for believing that there are possible worlds) somehow goes against my assertion that the possible worlds I believe in are not creatures of my imagination?

So much for the creativity – now what about the license? I take it that if you want to advance metaphysical theories at all, not just sit back and play the skeptic, then it is a good idea to try to find credible theories. I take it also that if you want a credible theory, you'd better try to avoid conflict with your own firmly held naive opinions, so far as the countervailing desiderata of strength and explanatory power and the requirement of consistency permit it. Call your opinions "linguistic intuitions" if you'd like to lend them some extra authority – personally, I'd rather not. Whatever we call them, the method of theorizing that balances them against systematic power and consistency seems pretty generally practiced, and I see no need to apologize for following it myself.

Must we then make consistent sense of every (grammatical) piece of metaphysical talk? Could we not regain naive set theory, for instance, by following my preaching and my example? Not quite. What we could regain is a defeasible presumption in favor of naive set theory. Defeasible – and now well and truly defeated. But worth regaining nonetheless. I think it is a real error to claim that iterative set theory gives us all we ever really wanted. It wasn't just *Frege's* naive set theory that fell; it was *ours*. What makes the paradoxes serious is that it is so very easy to recognize Frege's theory as

a codification of what we half-believed all along. Certainly we were wrong. But I don't think we understand yet what our mistake was. Certainly tria1-and-error ways of keeping out of trouble are a lot better than nothing, but I don't think we should lose sight of the problem that is still with us.

How do we know that possible worlds, or propositions for that matter, form a set? I don't suppose we have conclusive reasons to think they do or they don't. Scott mentioned proofs by diagonalization, and those, if produced, might be conclusive. It would presumably depend on the vulnerability of their premises. But let's cross that bridge if we come to it.

An inconclusive reason that has influenced me is as follows. There are unproblematic mathematical entities that can be regarded as representing distributions of matter in space–time, and thereby as representing the possible worlds where matter is so distributed. An account is to be found in Quine's "Propositional Objects." According to classical mathematics, these representatives do form a set. If every difference between worlds was reflected in a difference between their representatives, and if no worlds were unrepresented, then the worlds would correspond one-to-one with their representatives, which means that the worlds would themselves form a set. Now I don't suppose, and neither does Quine, that his system of representation is complete in either of the required ways. But there's enormous room for generalizing and enriching the system, and I do find it plausible as a working hypothesis that some or other improved version would do the trick. I don't know which improved version, exactly, and I don't even know that there's any way to find out which; but that's no argument that none will do. The field of candidates is very rich.

If the worlds do form a set, in the image of the set of their appropriate representatives, then so also do the sets of worlds. These I have called the propositions, conceiving of propositions simply as regions of logical space. But propositions are sometimes conceived differently, as language-independent entities that nevertheless have something analogous to syntactic structure. I would indeed question whether there is a set of all propositions under this second conception. It would be natural to suppose, for instance, that for any set *S* there is a proposition to the effect that the set *S* is self-identical; and that for any two sets the two such propositions are different. Then there is no set of all propositions. There may or may not be a problem here for a theorist who both wants to reduce worlds to books of propositions and prefers the second conception of propositions to the first. I don't see that there's a problem here for me.

I have long intended my realism about possible worlds to be a close imitation of commonplace realism about the entities of classical mathematics. Scott has suggested, and I think he may very well be right, that the two realisms are linked by more than imitation. I would be very interested to know whether the two must stand or fall together; but if so, I do not therefore conclude that they fall.

Some of Scott's further misgivings concern the idea of *comparative overall similarity* between worlds or between their inhabitants (his further misgivings about the uses I make of this idea will not concern us tonight). I usually explain overall similarity as a resultant, determined by the balance of very many similarities and dissimilarities in many respects, of which some weigh more heavily than others. It will be instructive to ask how much he is worried about the balancing, and how much he is rather worried about the host of similarities and dissimilarities in various respects that get balanced.

One way to worry about the balancing is to wonder how determinate are the relative weights of the various respects of comparison. I have myself insisted that they are determinate only within a fuzzy range, that this fuzzy range itself is not permanently fixed, and that any further determinacy would be arbitrary. I have insisted on this not by way of concession, but rather to claim an advantage of my approach over rivals that deny themselves the means to explain the observed shiftiness and indeterminacy of counterfactuals and *de re* modality. But I don't think it is the indeterminacy of the weights that bothers Scott; his worry seems to be not about selecting among an overabundance of similarity relations but rather about supposing there is even one. Perhaps the trouble is rather this.

Suppose, let us say, that there are countably many independent respects of similarity that each contribute to similarity with a weight of +1; and countably many respects of dissimilarity that each contribute with a weight of -1. The case confronts us with an insoluble problem in infinite arithmetic. Fair enough; I suppose such cases might arise, and might contribute to the indeterminacy and arbitrariness of comparative similarity. Especially so, perhaps, when we look at far-fetched possibilities, comparing two rival counterparts both very different from the original in very different directions; and, sure enough, questions about especially far-fetched possibilities do seem especially indeterminate and arbitrary. But I think Scott may think that such cases are all-pervasive, and I don't see why he should think so. I can certainly think of ways of weighting the respects of comparison that do make the problem all-pervasive, but that just shows that those ways aren't the right ones.

Maybe instead, or maybe also, Scott is worried about the respects of comparison themselves. I do think of possible worlds as unblurred and fully detailed; except in some very special cases, I take it that it would be an infinite task to describe a single possible world in full. So what? Scott asks *how we know that the world can be grasped in complete detail*. I suppose I know for sure that it *can't* be, at least if grasping is supposed to be something like having a fully detailed mental image or verbal description. Possible worlds are not meant to be mental representations, and to *be* is not to be *grasped*. If hypothesizing were creating, and if worlds couldn't be created many at a time, then I suppose only the graspable ones could get created. But hypothesizing isn't creating.

Note

1 Given at the Philosophical Society, Oxford, May 15, 1975. Scanned, tidied up, and conformed to the original typescript by Stephanie Lewis, August 23, 2003.

Part II

Methodology and Context

4

Lewis's Philosophical Method

DANIEL NOLAN

Lewis is famous as a contemporary philosophical system-builder. The most obvious way his philosophy exhibited a system was in its content: Lewis's metaphysics, for example, provided answers to many metaphysical puzzles in an integrated way, and there are illuminating connections to be drawn between his general metaphysical views and, for example, his various views about the mind and its place in nature.

A case can be made that Lewis's philosophy also exhibited a systematic methodological approach. I doubt that much of this was self-conscious on Lewis's part, at least at first: my conjecture is that his famous claim "I would have liked to have been a piecemeal, unsystematic philosopher, offering independent proposals on a variety of topics" (Lewis 1983, ix) would have applied to what he would have liked in the matter of philosophical method: for each question, adopt the method best suited to make progress with it, without requiring that the method necessarily be the same for different philosophical topics.

Even in his earliest philosophical writing, Lewis would not have been an anarchist about appropriate philosophical method. Clarity in presentation seems to have always been something he valued. Lewis, like most analytic philosophers (and probably most philosophers in general) preferred his deductive arguments to be valid rather than invalid. Lewis's work, even from the beginning, tended to proceed through "armchair" methods: by and large, Lewis did not aim to establish conclusions on the basis of detailed empirical investigations using, for example, statistical methods, or complex scientific equipment. It is not that he was averse to citing results of empirical studies in the natural and social sciences: Lewis (1989, 53) cites several discussions of US and Soviet weapons and intentions, for example, even though in the same paper he says "[a]s a philosopher, my business is with the coherence of positions and the range of logical possibilities – not with the truth of empirical hypotheses." Lewis (2004) is a paper about "many worlds" interpretations of quantum mechanics, which, while theoretical, is not obviously more so than many papers in theoretical physics, and like papers in theoretical physics is indirectly constrained by scientific discovery. Lewis did not himself publish the results of empirical investigations (Lewis 1995 is an exception) but, despite remarks like the one

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

quoted above, he did not see philosophy as concerned purely with a realm of facts inaccessible to empirical investigation.

As Lewis's work developed, his methodological views were articulated more explicitly as well. In the introduction to Lewis (1983) where he provides an overview of some of the strands running through his philosophy to that time, he presents (pp. x–xi) an overview of his philosophical method (as of 1981, at any rate). Topics raised in that overview will be the basis of sections 4.1 and 4.3 of this chapter, though I will also discuss his distinctive method of philosophical analysis, which is one of the most striking methodological features of his work despite not being mentioned in Lewis (1983, x–xi).

My intention in this chapter is not just to present what seem to me some of the more important methodological themes in Lewis's work, but to offer some of my own suggestions about how Lewis's own methodological stances can be improved. While doing so might risk arriving at methodological recommendations worse than the recommendation to do as Lewis did, it does follow his methodological practice in at least one respect. Lewis would never uncritically recommend a philosopher's philosophical views just on the basis that it was what so-and-so believed: and even though I depart from Lewis's methods in a number of ways, paradoxically this lack of *uncritical* respect seems to me one aspect of Lewis's technique to follow.

This chapter will be divided into three sections. I will begin with "starting points": places where Lewis looked for data for philosophical inquiries. This will be followed by a section on one of the distinctive features of Lewis's technique of philosophical analysis: the project of specifying theoretical roles and then identifying deservers of those roles. (This is one of the best-known pieces of Lewisian method, despite an absence of any discussion of it in Lewis's 1983 remarks on method.) Finally, I will pay attention to the project of weighing costs and benefits, and philosophical theory-selection in this way. I have also written about Lewis's philosophical method in chapter 9 of Nolan (2005): while I will touch on a number of the same themes, I have tried not to repeat material from there more than necessary.

4.1 Starting Points: Science and Common Sense

From Lewis's earliest published articles, Lewis expressed his conviction that science, particularly physics, is a very good way of finding out about the world. In 1966 he said, writing as a philosophical materialist, "[a] confidence in the explanatory adequacy of physics is a vital part, but not the whole, of any full-blooded materialism" (Lewis 1983, 105). In 1994 he expressed a willingness to "optimistically extrapolate the triumph of physics hitherto" to embrace a materialism according to which all the fundamental features of our world are physical (Lewis 1999, 292). Lewis also employed scientific premises in a number of his philosophical explorations: the existence of single-case objective chances is demonstrated by radioactive decay (Lewis 1986a, xvi, 83). One proof that not all explanation cites a cause of the phenomenon is provided by the explanation of why some collapsing stars (e.g. those that become white dwarfs) stop collapsing: Pauli's exclusion principle ensures "it's gone as far as it can go" (Lewis 1986a, 222).[1] Lewis (2004) is an extended discussion of what lessons we should draw from Everettian interpretations of quantum mechanics, should they turn out to be best supported by our evidence and theoretical considerations. Lewis was not a philosopher who saw his conclusions as insulated from the findings of science.

Despite this, Lewis has gained somewhat of a reputation for ignoring science, particularly physics, in his metaphysics. (One recent representative quote describes Lewis as "a philosopher who is often pilloried for his lack of engagement with science" (French and McKenzie 2012, 44).) There are a number of reasons some philosophers of physics in particular have an animus against Lewis, but

one of the main ones must be his notorious remarks about what quantum mechanics can teach metaphysics:

> I am not ready to take lessons in ontology from quantum mechanics as it now is. First I must see how it looks when it is purified of instrumentalist frivolity, and dares to say something not just about pointer readings but about the constitution of the world; and when it is purified of doublethinking deviant logic; and – most of all – when it is purified of supernatural tales about the power of the observant mind to make things jump. If, after all that, it still teaches nonlocality, I shall submit willingly to the best of authority. (Lewis 1986a, xi)

Lewis's description of physics (appropriately purified) as "the best of authority" has not mollified all physics-friendly readers of this passage. It can be read as dismissive of physics, but in my view a better reading of it is as a (polemic) statement of a preferred philosophical interpretation of quantum mechanics; or at least the view that the best interpretations are not instrumentalist, quantum-logical, or Copenhagen-esque. Partisans of those interpretations might object, but I do not think they should object to this as a case of anti-scientism: rather, it is a case of Lewis taking sides in a live dispute in theoretical physics (a dispute that may seem a little dated now, but those remarks were written in 1984, albeit published in 1986). Respecting the deliverances of science need not require not taking any position in scientific debates themselves: it would be untenable to believe everything scientists say when scientists themselves disagree. (Nor is it clear that we should always trust scientists as having the last word about what is in fact shown by scientific evidence, though of course specialists will often be best placed to make those judgments.)

The other important starting point for Lewis is what I will label "common sense." Lewis thought that our ordinary starting opinions were important constraints on our philosophical theorizing, and that straying too far from common sense is methodologically out of bounds. A theory "cannot gain, and it cannot deserve, credence if it disagrees with too much of what we thought before. And much of what we thought before was just common sense" (Lewis 1986b, 134). Common sense starting points are all over Lewis's philosophy: counterexamples to rival analyses are often common sense ones, folk psychology is an important starting point for his philosophy of mind, his theories of causation rely on common sense observations as well as more unusual thought-experiment cases, and so on.

Lewis's respect for common sense and starting opinions takes two major forms. One is that we cannot diverge too far *overall* from "what we thought before": a generalized theoretical conservatism. The other is a more specific attitude to some pieces of common sense: the "Moorean" attitude about some *particular* claims that they are more certain than the premises of any argument to the contrary could be (or perhaps any *philosophical* argument to the contrary).

Lewis uses the expression "Moorean facts" to characterize these particular undeniable claims, and at one point says we "lose our Moorings" if we deny the existence of sensations, simultaneity, and values (Lewis 1999, 418). In doing so, Lewis is alluding to G.E. Moore's famous claim that certain obvious truths (such as the fact there were two hands in front of him) were more obvious than any philosophical premises that yielded conclusions to the contrary. Furthermore, Moore thought, our knowledge of such obvious truths was sufficiently sure to refute skepticism: a theory of knowledge that entailed Moore could not know there were hands in front of us is refuted, since he knew there were hands in front of him (Moore 1959).

There are a number of different things that Lewis counts as "Moorean facts" in different places in his work, and many of these are parts of apparently secure common sense. That there is "apparent sameness of type," for example that things sometimes are the same shape as other things, is a Moorean fact that any theory of properties and relations should respect (Lewis 1999, 20). That many of our ordinary beliefs about colors are close to true (Lewis 1999, 333) is Moorean. That our

language has a fairly determinate interpretation (Lewis 1999, 47) is Moorean. That we know a lot (Lewis 1999, 418). Many of Lewis's philosophical investigations start from starting points like these: not negotiable pieces of our ordinary picture of the world.

Why think any of our beliefs have this special status of not being up for grabs? Lewis never says explicitly, and a number of his characterizations of Moorean facts are ones that rightly apply to *any* known facts (see Nolan 2005, 208). Perhaps it is because of a limited faith in the powers of philosophy; Lewis at one point makes fun of those who would use philosophy to overturn apparently secure knowledge: in this case, those who would rely on philosophical arguments that standard mathematics is full of falsehoods.

> If they challenge your credentials, will you boast of philosophy's other great discoveries: that motion is impossible, that a Being than which no greater can be conceived cannot be conceived not to exist, that it is unthinkable that anything exists outside the mind, that time is unreal, that no theory has ever been made probable by the evidence (but on the other hand that an empirically ideal theory cannot possibly be false), that it is a wide-open scientific question whether anyone has ever believed anything, and so on, and on, *ad nauseum*? Not me! (Lewis 1991, 58–9)

Lewis did not think these "discoveries" were genuine, of course, but his point presumably is that philosophical argument has often led us astray, and badly astray at that. A look at the track record of philosophical argument should make us dubious that it can be used to overthrow apparently very secure common sense opinions. (Or secure deliverances from sciences like mathematics, for that matter.) Or so Lewis's suggestion seems to go.

Some general confidence in our epistemological abilities seems reasonable, especially when beliefs seem to be part of generally successful epistemic projects. I am a reasonably good detector of color, of tables and chairs, of whether the Earth is round, or whether plants are alive, among a host of other questions. To the extent I think I know many such things, I should believe that it will not turn out that I was wrong. Perhaps I even should have an expectation, about each thing I think I know, that I will not turn out to be wrong about *that*. But Moorean confidence seems stronger than this. It seems to be strong enough to put one is a position to dismiss any arguments or evidence to the contrary. Many of the things I know are not like this: I know that none of my colleagues are Kierkegaard specialists, for example, but I could be convinced otherwise (perhaps merely on people's say-so, perhaps by a bit of internet searching that discovers my colleague's seminal books on Kierkegaard).

I doubt that there is anything very interesting that is so secure that *any* argument to the contrary is rightly dismissed by mere modus tollens. Those that think that philosophy is a particularly weak source of reasons to believe might think that, at any rate, there is a body of ordinary knowledge that should be immune to *philosophical* challenge (it would be "presumptuous," to use Lewis's expression for the envisaged philosopher who disagrees with mathematicians). But philosophers do not seem particularly limited in the methods we use: philosophers who want to challenge our ordinary color judgments in part by appeal to color science, or philosophers who want to challenge our ordinary temporal judgments by appeal to cutting-edge physics, do not seem to be doing anything outside the bounds of philosophy. So if a challenge can be launched at all against apparently well-grounded doctrines in common sense or science, then it is hard to see how philosophers could not legitimately launch such challenges.

While Lewis's view that there are near-unchallengeable Moorean facts is distinctive and clearly relevant to his method, my view is that not much would need to be done differently in pursuing Lewis's own projects, or Lewis-style philosophical projects, if this commitment to Moorean facts was watered down. Even if no claim was immune to philosophical challenge, still we might be reluctant to challenge beliefs that we held strongly and which were apparently part of successful common sense

pictures of the world. Even if we thought that in principle we could give up the view that there are colors in the world and they are sometimes properties of surfaces of objects, it would certainly make sense to *first* see if there was a satisfactory theory of color that preserved that opinion. One could think that it counts very heavily against a theory of knowledge that it cannot allow, for example, that I know that have hands, without needing to insist that "I know that I have hands" is a non-negotiable constraint on any theory of knowledge. Giving common sense significant weight but not maximum weight will make it secure enough to allow us to choose between philosophical theories on the basis of how well they respect key pieces of common sense, whether or not we count any of common sense as Moorean.

The methodological question of whether we should give any part of common sense a lot of weight in our philosophizing remains, even if we do not go as far as Lewisian Mooreanism. I am inclined to think that we should, at least in the early stages of our inquiry, and at least for parts of common sense that appear to be well supported. (We should be more confident in the existence of cats than that extraterrestrials will seem weird to us, even if both are "common sense.") Not to do so would be as strange as starting geology without the assumption that there are rock formations, or chemistry without the assumption that fire can change substances put in it. (The call to ignore ordinary opinion in philosophy often seems to come from those who want to make philosophy more like the sciences, but I do not think a reconstruction of the epistemic position of the sciences is adequate if we ignore the common sense bases of those sciences' starting points.) That is not to say that we might not eventually give up much of (current) common sense, but it is not clear that specifically philosophical investigation is in a good enough position to do anything like that yet. Still, the role of common sense in philosophy and inquiry generally is too large an issue to try to resolve here.

One reason not to privilege common sense *per se* is that many of my other beliefs also seem very well supported and resistant to change on philosophical grounds: why should it matter very much whether such beliefs are *common*? Perhaps we had better take our ordinary opinions as constraints on our theorizing, not so much because of any special status they have in themselves, but simply because they are among our initial beliefs: perhaps it is *our initial beliefs* that ought to have a special role to play in philosophical inquiry. "Theoretical conservatism" is an expression that is used for a variety of approaches, but they share the view that one ought to continue to hold one's starting beliefs until one has good reason to change them. (Whether this is just because they are the starting beliefs, or whether it is because there is reason to suppose one's starting beliefs are in the main warranted, is one matter that defenders of conservatism may disagree on.) Lewis thought that we had no reasonable choice but to respect our starting beliefs:

> It's not that the folk know in their blood what the high falutin' philosophers may forget. And it's not that common sense speaks with some infallible faculty of "intuition". It's just that theoretical conservatism is the only sensible policy for theorists of limited powers, who are duly modest about what they could accomplish after a fresh start. Part of this conservatism is a reluctance to accept theories that fly in the face of common sense. But it's a matter of balance and judgement. (Lewis 1986b, 134)

Despite the reference to common sense, the "only sensible policy" here seems to be to not lightly tamper with one's starting point: to be modest about what could be accomplished after a fresh start. Lewis never explicitly discusses what one ought to do if one were to come to believe that one's own starting beliefs were idiosyncratic and not "common sense," but theoretical conservatism would likely take as dim a view of a wholesale revision to match the beliefs of others as any other wholesale revision.

The version of theoretical conservatism that Lewis advocates appears to be one that does not allow us to move beyond a certain point from our starting opinions. He says:

> What credence [a theory] cannot earn, it must inherit. It is far beyond our power to weave a brand new fabric of adequate theory *ex nihilo*, so we must perforce conserve the one we've got. A worthwhile theory must be credible, and a credible theory must be conservative. It cannot gain, and it cannot deserve, credence if it disagrees with too much of what we thought before. And much of what we thought before was just commonsense.

It is not entirely obvious when "before" is here, but it seems to me to be something like "before we started theorizing." That is when we can be most certain that most of what we thought was common sense. Notice again that it seems to be common sense's role as our starting opinion that makes it fit for belief here: this passage again suggests that it is common sense's role as our initial opinion, plus the imperative to conserve what we have, that explains why philosophical views should not depart too far from common sense.

Lewis had another particular reason to embrace conservatism, given his other philosophical commitments. Lewis's favorite story about belief revision was a Bayesian one, according to which an ideally rational agent starts with a distribution of probabilities over all the possible contents of belief: these are the agent's rational degrees of belief in each of those options (and in the case of the entirely rational agent, her rational degrees of belief will match the degrees of belief she has as a psychological matter). Equipped with these degrees of belief, the agent then *updates* on new information that comes in, for example, by the senses. Lewis's preferred account of updating, for the entirely rational agent, is by *conditionalizing* (Lewis 1999, 403–7). The agent learns some perceptual information with certainty (probability 1), and then her other degrees of belief change in light of the information learned: when she learns A, the new subjective degree of belief she assigns to a proposition B is equal to the old value her belief system applied to the conditional probability of B on A ($P(B|A)$).

Lewis was prepared to concede that agents of more limited powers than the ideally rational agent might legitimately update with a less demanding rule, for example, that of Jeffrey conditionalizing (Jeffrey 1983): in effect, limited agents like us may be legitimately uncertain of what we have learned, and so not give any direct deliverance of our senses a probability of 1 on a given occasion. And presumably limited agents like us depart from ideal rationality in other ways as well: perhaps our beliefs are sometimes probabilistically inconsistent from time to time (e.g. when we believe a number of propositions separately that are jointly inconsistent: see Lewis 1998, 97–110). Across time, our degrees of belief might sometimes shift without relevant evidence, as when we become aware of new hypotheses or are affected by wishful thinking or suffer imperfect recollection of our past opinions. Even when we accept that the theory of the rational Bayesian agent is an imperfect model of what beings with our powers can realistically aspire to, one feature of it in particular suggests that even beings like us ought to be theoretically conservative.

It is part of the standard Bayesian model that there is a fair amount of leeway allowed in the starting rational degrees of belief of an agent (their "prior credences"). Agents will start out almost certain of some contingent propositions, for example (especially when a model has infinitely many propositions in it). And typical models of Bayesian agents over time explain the rationality of later states in terms of how the agent got there from earlier states: for a defender of conditonalization like Lewis, the agent's credences *now* have to be produced by conditionalization from her states *before*, if there is a change. Being automatically rational in having initial opinionated states not derived from anything, together with the rationality of later states being largely derived from how they relate to states had before, suggests a picture where the mere having of the earlier states, together with the evidence that has come in as input, justify the later states. (The core of Bayesianism talks of rational degrees of belief rather than "justification," but this is one obvious way to try to map talk of justification into a Bayesian framework.)

Admittedly, a Bayesian formal framework is compatible with shifts in credence very unlike what we would expect conservatism to endorse. An iconoclast who radically changes most of her unconditional degrees of belief each time she receives a piece of evidence can be modeled without violating any formal constraints (the conditional probability of B on A can be wildly different from the unconditional probability of B). But the spirit that motivates Bayesianism can motivate more conservatism than the letter of Bayesianism entails. *Plausible* conditional probabilities of ordinary sensory input will often leave much of what is believed largely intact: I see my cup on the desk, and that should normally not radically change my opinion about the Mongol empire or the existence of isobutane. If this hunch about the rational conditional degrees of belief of creatures like us is correct, then we get a picture where my degrees of belief now are not only rationally determined by my degrees of belief before, but that, rationally, my unconditional beliefs will tend to remain largely unchanged until evidence impacts on them relatively directly. If this is the right picture of rational agents, it suggests the rational agent will be theoretically conservative.

My suspicion is that this is not the only point of contact between Lewis's broadly Bayesian views about rational belief revision, on the one hand, and his views about philosophical method (and appropriate belief revision in philosophy), on the other. Another obvious point of contact is between, on the one hand, the Bayesian sympathy for equally rational prior probabilities yielding different but equally rational reactions to evidence, and on the other hand, Lewis's conviction that rational philosophers presented with the same philosophical considerations can rationally disagree about their conclusions (see section 4.3). Since Lewis never explicitly connected the topics of Bayesianism and philosophical method, drawing connections requires a certain amount of reconstruction, and so the task of drawing out other links is best left to another occasion.

I also have some sympathies for theoretical conservatism, and agree that there is a danger in philosophy of too quickly shifting to radical or exciting views that abandon too much of what we thought before on a topic. (Even those who oppose philosophical conservatism should think this is *one* of the dangers, just as even philosophical reactionaries should concede that *one* of the dangers is not shifting one's philosophical position enough in the face of arguments and evidence.) Lewis's version seems to me unduly strong. Contrast Lewis's view, that we cannot (rationally) end up disagreeing with our starting point beyond some limit, with *stepwise* conservatism. According to stepwise conservatism, each transition we make cannot depart too far from our initial network of beliefs. But once we have made some rational transitions because of evidence or argument, when assessing whether a further transition is rational we consult our *current* opinions, rather than evaluating the envisaged transition partly on the basis of where we began our theorizing. Maybe a new theory would have struck us as incomprehensibly radical 20 years ago. But if it seems like a measured response to our evidence and argument now, why should it worry us if "it disagrees with too much of what we thought" 20 years ago?

If one is a Lewisian conservative (as I have interpreted his view, at least) rather than a stepwise conservative, one faces an awkward challenge of explaining opinion change through time *across* individuals. Suppose there is some set maximum "doxastic distance" which someone can reasonably travel before one "disagrees too much" with what they thought before. Adam begins rational enquiry with beliefs B_1, and after a life of unexpected surprises and philosophical reflection rationally ends up with beliefs B_2, near the edge of the maximum distance from B_1. Adam raises his daughter, Belle, with a B_2 world-view, and by the end of her life of inquiry she (rationally) reaches B_3, near the maximum distance from B_2 and almost twice the maximum distance from B_1. If this is coherent, a puzzle arises.

Consider Adam's very long lived brother Carl. He also begins at B_1, lives the life of the mind with Adam until he dies, and then follows Belle's fascinating work thereafter. Presented with the same lines of argument and evidence as Adam and then Belle, he also ends up at B_3. But why should we

have to judge Carl's overall trajectory irrational when Adam's and Belle's were both rational? Why would Belle, but not Carl, be allowed to take note of Belle's excellent evidence and argument to the conclusion B_3 from their shared starting point B_2? If we erased Carl's beliefs around the time of Adam's death and "restarted" him at B_2, then would he have been allowed to revise onwards to B_3? Rejecting stepwise conservatism seems to leave us having to make an invidious and undermotivated distinction between Carl and Belle once Belle starts questioning their shared B_2 views.

There are various ways a Lewisian conservative could resist this case if she wished to say that Belle and Carl are restricted to roughly the same theoretical options when Belle begins her independent inquiry. Perhaps I have been unreasonable in thinking that anything like B_2 is a fit place to *begin* inquiry. If we all started in roughly similar "pre-theoretic" states of nature, for example, Adam might already have had to move Belle a long way from her starting point to get her to B_2. Or perhaps Belle is only permitted to move so far from B_1 because she *lacks* a lot of reasons and evidence that those who have been to B_1 possess; maybe there must be some epistemic *advantage* Adam and Carl have that she lacks (despite holding Adam and Carl's views at the beginning of her inquiry), and it is this advantage Carl has that makes it irrational to follow Belle's course. Here is perhaps not the place to consider every feasible response a Lewisian conservative might offer, though any of these would need significant fleshing out if we are to be convinced that Belle may go where Carl cannot after they both endorse B_2.

There are other uncomfortable questions that face the Lewisian conservative. We unfreeze Oog from the ice she has been trapped in since the Pleistocene. How close may we bring her to contemporary opinions in physics, economics, or philosophy through evidence and argument, if we want her to be able to rationally believe things we know? Abstracting from practical limitations she might have, we might imagine that eventually we can teach her the latest about fundamental physics or economics or philosophical methodology even though the starting point for her enquiry might be considerably further back than those of us brought up with contemporary assumptions about how the world works. The stepwise conservative, on the other hand, may need to concede that Oog may not be able to jump straight from flint-knapping to quantum mechanics or from tribal ceremonies to the theory of the liberal state, but can at least hold out the prospect that in principle she can learn the things we know by stages, maybe via imperfect approximations of what we take to be the truth (Newtonian physics, Locke on the state, etc.).

Even stepwise conservatism requires defense, of course: one might even think that it is rationally permitted to go through a "conversion experience" and move all of one's views at once to something radically different (see van Fraassen 1989, chapter 7). Unfortunately a discussion of theoretical conservatism quickly raises many of the most fundamental issues in epistemology and belief revision. Whether supported by good arguments for theoretical conservatism or not, Lewis adopts common sense as a starting point as well as the deliverances of mature natural sciences. But the starting point is not the end point: armed with the starting points, we can turn to trying to solve philosophical problems.

4.2 After the Starting Points: Defining Theoretical Roles, Finding Deservers

A particularly influential part of Lewis's method concerns a way of approaching philosophical puzzles. Questions as diverse as "what is pain?," "what is meaning?," "what are properties?," and "what is moral value?" were all, at one time or another, treated in the same general way. One aspect of this approach is now called the "Ramsey/Carnap/Lewis treatment of theoretical terms," and another might be called a kind of "generalized functionalism." It has been the inspiration for philosophical movements like the "Canberra Plan" (see Braddon-Mitchell and Nola 2009), and its

influence can be seen in scores or hundreds of papers in the current philosophical literature. Since this method has been extensively discussed elsewhere, including in another chapter of this volume (chapter 5; see also chapter 33, section 33.2), a brief sketch here together with some critical remarks might suffice.[2]

First, one assembles a theory featuring a term apparently referring to the phenomenon of interest. It may be as easy as taking a canonical theory from a science (as we might do if we were interested in "what are electrons?" or "what is an ecosystem?") or, if the term is one in widespread unsystematic use, we might need to articulate a "folk" theory of the phenomenon of interest. For example, if we want to define pain or belief, we are to use "common sense psychology" as the relevant theory (Lewis 1999, 249). Then one uses that theory to define a "role" associated with the phenomenon. If the expression in the theory we are interested in is a predicate, we first transform the theory to talk about the associated property. (Instead of "when people are in pain they tend to cry out" we might use "when people have the property *pain* they tend to cry out," for example.) Then, when the theory is modified so that the expressions of interest are all nouns, we can create a "matrix" by replacing the terms of interest with variables, a different variable associated with each term of interest ("when people have *x* they tend to cry out").

One thing we can do here is use a theory to simultaneously provide a matrix for a number of different terms and so provide a role for a number of things at once: the role of belief might be specified partly in terms of the role of desire and the role of intention, if it is part of our folk theory that beliefs and desires go together to produce intentions. The aim is to provide a matrix where a range of puzzling vocabulary is replaced with variables, but enough else is said using expressions that have not been replaced to give the matrix substantial content. (We might offer a matrix replacing *all* the psychological terms in a theory of mental states, for example, leaving only causal vocabulary and specifications of non-mental inputs and outputs, such as perceived objects on the input side and behavior on the output side.)

Armed with this matrix, we can treat it as being associated with a *role*. That *role* is satisfied by a collection of entities (properties, events, "things" or whatever) when those entities collectively satisfy the matrix. (To use the same kind of example Lewis uses (in Lewis 1999 249–52), if our matrix is "X killed Y in the kitchen of the mansion using Z," and Anne killed Bertrand in the kitchen with Excalibur, then Anne, Bertrand, and Excalibur collectively *satisfy* our matrix.) We call the entities that jointly satisfy the matrix the *realizers* of the role associated with the matrix.

Lewis did not think that an expression had to pick out something that satisfied *every* clause of such matrices: even canonical theories might be wrong in some detail. To count as the property *pain*, or the property *red*, or the property *morally right action* the relevant properties have to play *enough* of the relevant role, and to play more than any rival properties. In early applications, Lewis emphasized the construction of *causal* roles: roles specified largely in what the typical causes and effects of entities and states are. While that seems appropriate for some philosophical projects, it does not seem crucial to the method, and applications of the method to , for example, mathematical objects, or possibilities, or values, seems a feasible thing to try even if we doubt the roles taken from available theories will have much to do with what sets or possibilities or values *cause*.

The methodological point of extracting roles from theories and talking about potential realizers that satisfy all, or most, of those roles, can sometimes be a little obscured by the framework of variables, matrices, definitions and so on (particularly because usual presentations often talk about "Ramsey sentences," "Carnap sentences," strings of quantifiers, and other technical devices). One methodological purpose of this focus on roles and what realize them is to try to extract *definitions* that can provide analytic truths about the topics of interest. For example, we can attempt to define "pain" as that which best satisfies enough of the pain role, "value" as that which best satisfies enough of the value role, and so on. Providing definitions and an account of the analytic/synthetic

distinction was one of Carnap's main interests in using this sort of machinery (see Carnap 1963, 958–66). Using this method to extract plausible definitions is not entirely straightforward, however, especially once the issue of deciding which theory is canonical is taken seriously (Nolan 2005, 219–22, expresses some preliminary worries). However, I think another aspect of employing this method is where the more important methodological implications lie.

Constructing a role from an entire theory and looking for entities that play that role tends to do two things. One is that it shifts the focus from trying to isolate a crucial core to a philosophical concept to paying attention to many theoretical connections: instead of trying to work out what the special mark of belief is, for example, we look for what it is that is connected in the right way to perception, desire, action, and so on. This style of extracting a characterization of something from an entire theory of it is sometimes called "network analysis" (e.g. Smith 1994, 44–56) because of this feature and because of the way it can be applied to a family of puzzling phenomena at once: all of the mental in terms of causal relations with the physical, or all of the semantic with linguistic behavior, for example. Instead of trying to isolate a balance between necessary and sufficient conditions, for example, assembling many indicators and looking for what satisfies most of them allows a theorist to not have to immediately defend some particular criterion or two as all-or-nothing features of the phenomenon under discussion.

The second tendency of this way of approaching philosophical analysis is that, in the first instance, it focuses less on the *natures* of entities than the relationships between those entities and other phenomena (hopefully some of which are better understood). For example, instead of trying to reflect directly on the nature of moral value, Lewis's style of analysis encourages a theorist to articulate connections between moral values and other matters: rational behavior, desires, moral obligation, and so on. The focus is more on what something of interest *does* than, in the first instance, on what it *is*. This might be literally true, when an analysis in terms of a causal network is offered (for beliefs, for example, or for colors), or more metaphorically true (when giving an account of propositions, or of moral values, or of numbers), when the theoretical role is unlikely to feature much about causation.

Both of these shifts have been found useful and liberating by philosophical inquirers. Another advantage, it seems to me, is that questions of which part of the theory are necessary and which contingent, and questions of which are conceptual truths and which mere synthetic ones, do not need to take center stage in this style of philosophical analysis. As long as we have a *true* theory concerning an entity of philosophical interest, our target will be among the potential realizers of the role we extract from that theory, and if it is plausible that only one thing plays the relevant role, then it will be plausible that that is, indeed, our target. If I tell you a cluster of truths about the (surface) color scarlet, and a particular physical property of surfaces is the unique property that satisfies the role defined by those truths, then that property must be scarlet, even if the truths we used to define the role were contingent and synthetic. After all, if only one thing does what the color scarlet does, it is the color scarlet.[3] Many who deploy the Lewisian method of analysis hope to establish analytic or conceptual truths with it, and the method does not preclude doing so, but those inclined to leave the issues of analyticity and conceptual truth to one side can still employ the method to derive interesting philosophical conclusions.

It should be conceded that Lewis's suggestion about philosophical analysis is far from the only way to gain the benefits of looking at many theoretical connections rather than few, and of looking at what an entity is supposed to do rather than, in the first instance, what the nature of that entity is. It should also be noted that one can offer something in the form of a Ramsey–Carnap–Lewis analysis, complete with theoretical role and identification of theoretical realizer, while engaging in a project with neither of these features. "x is justified, true belief" is a matrix we could associate with knowledge, if we wished, and then we could find a state that played that role (primitive knowledge,

for example, or beliefs that were both justified and true). Unless we went on to do something more substantive with roles for justification or truth or belief, that would not bring some new advantages to a traditional JTB (justified-true-belief) analysis of knowledge. Employing Lewis's method of philosophical analysis does not *necessarily* bring the features with it that have proved fruitful. But philosophers thinking about philosophical analysis in Lewisian terms are at least nudged in the direction of networks and functional roles, and this nudge has been enough to yield a recognizable family of philosophical theorizing. The results of this family, in the work both of Lewis and others, seem fruitful and exciting: at least to an admitted insider like me.

4.3 Counting the Costs

One important part of Lewis's philosophical method is that he saw deciding between philosophical questions as a matter of weighing up costs and benefits, and selecting the theory which, in the judgment of the weigher, did best by that measure. Lewis was not alone in this: many contemporary philosophers will pay at least lip service to the idea that what must be done when deciding a philosophical issue is to weigh up costs and benefits of rival philosophical views. Perhaps because this is such a widespread idiom, it seems worthwhile paying closer attention to the role this played in Lewis's philosophy. What is it to weigh up costs and benefits of a theory? What is a cost, what is a benefit, and how is the weighing to be done?

The first place this way of talking is prominent is in "Holes," a paper co-authored with Stephanie Lewis (Lewis and Lewis 1970). The paper is a dialogue between two characters neither of whom is a spokesperson for the authors, but where the characters agree it is probably safe to assume their views correspond with their authors'. There, the characters talk of "paying a price" of plausibility, when a theory disagrees with common opinion, and a theory "earning credence" through clarity and economy. The characters also suggest that many debates over "ontic parsimony" are a matter of counting the costs of disagreement with common opinion on the nominalist side, with the advantages accruing to a nominalist when his or her theory is economical and/or clear (Lewis and Lewis 1970, 211–12).

Lewis endorses this language of costs and benefits in his own voice in Lewis 1983. Even after we have a stock of counterexamples before us and have heard all the philosophical arguments, "presumably we will still face the question of which prices are worth paying, which theories are on balance credible, which are the unacceptably counterintuitive consequences and which are the acceptably counterintuitive ones" (p. x). This might sound like an ultimate appeal to intuitions for assessing theories, but it is not quite that: on Lewis's view, "[o]ur 'intuitions are simply opinions' " (p. x). It looks like this assessing of prices, and working out the credibility of theories, must be done by the lights of our opinions taken as a whole.

Lewis also uses the language of costs and benefits in Lewis (1986b, 135). He admitted that his concrete modal realism involved a "denial of common sense," and said "I think it is entirely right and proper to count that as a serious cost." Nevertheless, he thought, "the price is right" since "the theoretical benefits are worth it." As well as that judgment, though, he thought that he needed to show "that they cannot be had for less": concrete modal realism about possible worlds was better than rejecting possible worlds, but should only be believed if it was also sufficiently better than ersatz modal realist alternatives. In this passage Lewis also contrasts losses from disagreements with common sense with gains from theorizing, this time the "earned credence that is gained by making a theory more systematic" (Lewis 1986b, 134).

Enough is said in these passages to get a general sense of what Lewis has in mind. The "costs" and "benefits" seem to be in credibility, or gains and losses of credence theoretical options receive. A

theory can gain benefits from agreement with our earlier opinions, but there are other ways to gain benefits as well: clarity, economy, being systematic. I assume Lewis would agree that there are other ways to gain benefits as well: empirical support from scientific inquiry, for example.

One puzzling thing about some of this discussion are the things Lewis thinks can be "benefits," if the goal is to select the most belief-worthy theory. Why would clarity or being more systematic add to the credibility of a theory? Both can sound like mere matters of presentation. They might still play an important role for less than ideal theorizers – perhaps clarity can show some apparent problem to be merely apparent, and a systematic theory can make it obvious how evidence bears on hypotheses in a way that a jumble of evidence statements and theoretical assertions does not. But I suspect Lewis had something more substantial in mind. One way to "clarify" a theory is to stop running two phenomena together; and this can be particularly useful if you initially said inconsistent things. If A is F, and B is not-F, and A and B were initially confused, then instead of one thing described as F in one place and not-F in another, a theory that says there are two things, one F and one not-F, will be much more believable. "Clarification" need not be this extreme, of course, but drawing distinctions can be more valuable than merely providing convenience of presentation.

I suspect fuller-bodied theoretical virtues are relevant in other ways here as well. "Economy" of a theory could just be succinctness, but Lewis's remarks suggest he has in mind a virtue like parsimony of theoretical postulates. "Systematicity" might signal a unified theory, with theoretical postulates each confirmed by, and perhaps explaining, a range of evidence. (Lewis speaks of "trying to improve the unity and economy" of a theory (1986b, 134), which also suggests he values unification in theorizing.) Lewis never offers us a complete list of the features that confer theoretical "benefits" to a theory, nor, for that matter, do we ever get an exhaustive list of costs. I suspect it would not be easy to discover such lists. It could well be that the task of saying what bears on the costs and benefits of theories, and the weights that should be attached to those costs and benefits, is about as hard, or harder, than working out what is the best philosophical position on each philosophical issue. One might even doubt that there is a general answer available here, if one was enough of a methodological particularist, though there is no evidence that Lewis, at least, was a particularist in these matters. It is worth noting that Lewis thought these costs and benefits were relevant to the question of which philosophical theory to believe (or give credence to): so he seems, at least implicitly, to have been committed to the view that features like clarity, systematicity, and economy play a role in belief-worthiness, not merely practical aspects of theory acceptance and manipulation.

Another puzzling thing about Lewis's discussion is what he means by "credence" here. Philosophical conjectures are supposed to gain or lose "credence" as we discover clashes with intuition or pretheoretic belief, or come to appreciate their economy or other systematic virtues. Elsewhere (e.g. Lewis 1983, 83, 110; Lewis 1999, 404), Lewis uses "credence" to pick out a rational degree of belief, and sometimes this rationality is quite idealized. A suitably idealized agent gives probability 1 to every proposition true at all possible worlds, and so we might expect that our "credence," our idealized degree of belief, in any necessary truth is also 1. Insofar as a philosopher is considering a hypothesis that is necessarily true or necessarily false, that philosopher's "credence" in that proposition, if idealized enough, is already either 1 or 0, and would not change by noticing things like its agreement with common sense, its economical capturing of phenomena of interest, and so on.

Whatever Lewis means elsewhere, I think we should not take him to intend that "credence" is such an idealized matter when we discuss philosophical method. Lewis seems to be offering advice that is not merely descriptive – it is not that whatever we happen to do with philosophical arguments is automatically right – but is not idealized to the point where we already endorse every necessarily true philosophical hypothesis. Or at the very least, we do not endorse every necessarily true philosophical hypothesis *in the guise that it is presented*. Lewis does hold that there is only one necessary

proposition and we already believe it, in at least one good sense of the term "proposition." But, when he says this, he allows that we might be mistaken about exactly what our sentences express (see Lewis 1986b, 36, where he cites Stalnaker 1984 approvingly.) So when we are trying to work out which philosophical doctrine deserves our approval, it may be that sometimes Lewis would characterize what we are doing as trying to work out which sentences express the necessary proposition. (Indeed, this might also happen with contingent philosophical matters: we might in some sense know all the relevant facts, but be unsure which philosophical sentences correctly capture the phenomenon we are already familiar with.)

To sum up, once we have a range of options before us, preferably clarified and systematized, we have to weigh them up before coming to a verdict. In doing this weighing, a crucial role must be given to their agreement with starting opinion: it is a cost insofar as they depart from what we believed before (and, according to Lewis, a decisive cost if they vary from any "Moorean facts"). Consistency with scientific findings, particularly well-established scientific results, is also a plus. Theories must also be assessed for unity, economy of postulation, and other such virtues, and these can count for a lot: for example, Lewis concedes that his modal realist metaphysics amounts to a "severe" denial of common sense, but nevertheless thinks that the theoretical benefits it brings make it worth the cost (Lewis 1986b, 135). Incidentally, this is a case where we are plausibly choosing between hypotheses that are all necessary if true and impossible if false: it is not as if the question of whether there are a plurality of concrete possible worlds is one which has an answer that varies from possible world to possible world.

My own view, for what it is worth, is that assessment of philosophical theories for theoretical virtues such as simplicity, unificatory power, track record of apparently successful problem solving, and so on does, and should, play an important role in philosophical theory choice. (And not just philosophical theory choice – in choices among more theoretical hypotheses in general.) Lewis's use of evidence from common sense and the findings of science also seems to me right: attempts to find a completely disconnected method for specifically philosophical questions seem misguided, especially as our philosophical opinions need to mesh with our other opinions about the world. Despite Lewis's protestations that common sense has no "absolute" authority in philosophy, insisting that we can never move too far from our starting position seems an overly strong application of theoretical conservative principles: once we have used our starting points to make progress from, it should be an option that as we discover more, we eventually move ever further and further from our beginnings, whether those are the beliefs of individuals before systematic inquiry or the beliefs of our ancestors (Quine's *homo javenensis*). I also suspect that Lewis's scattered appeals to "clarity," "unity and economy," and so on risk downplaying the methodological heavy lifting that principles of simplicity and parsimony, unity, and perhaps other forms of theoretical support must play if we are to justify philosophical systems. And not just philosophical systems – even an explanation of how Newton's physics was justified given the evidence he had available will require heavy reliance on the epistemic role of theoretical considerations, it seems to me, let alone a justification for a contemporary overall scientifically, informed view of the natural world.

Lewis also suggests that there is a certain lack of decisiveness in philosophy as a result. "Philosophical theories are never refuted conclusively," Lewis says, though he adds "Or hardly ever" (Lewis 1983, x). Lewis thinks there are no "knockdown" arguments in philosophy, or hardly any, because in practice there is almost always some option open to someone who holds a view and who wishes to maintain it. Lewis might just be making a psychological point about philosophers here – that they are too bloody-minded to all be convinced by an argument – but I suspect he is suggesting in addition that there is almost always some way of holding on to an antecedently held view in the face of problems, if someone has an unusual enough calculus of costs and benefits, that is at least somewhat rational.

"Once the menu of well-worked out theories is before us, philosophy is a matter of opinion" (Lewis 1983, xi). It is unclear from Lewis's remarks whether he thinks this is a matter of what philosophical disagreement is like *in principle*, or whether, for example, eventually ideally rational philosophers in possession of all relevant evidence would converge. One reason to think that they may not converge is the Bayesian one mentioned above: if they rationally start from different places, and rationally allow the evidence to impact on them differently, there is no guarantee they will end up in the same place.[4] It probably does not matter very much which, methodologically speaking: if it is not feasible to secure rational agreement between careful, intelligent philosophers of goodwill possessed of the same considerations, we are likely to be stuck with these divergences, and with the absence of "knock-down" considerations in many cases, for the foreseeable future.

4.4 Conclusion

The methods of philosophers are often not very clearly defined compared to the methods of those pursuing archival work, or studying gene expression, or doing many of the other research tasks in contemporary disciplines. While this can make it hard for students to see how to do philosophy well, or to get consensus even among professional philosophers about the quality of different pieces of philosophical work, it is clear that not anything goes, and there are ways of investigating philosophical questions that are better than others. The challenge for philosophical methodologists is to say something useful about how philosophy is done well, and how to do it better, navigating between the one danger of being too trite and the other danger of proposing a theory of philosophical method that is innovative but wrong-headed.

The recommendation to follow the example of philosophers who seem to be doing philosophy well seems like one piece of advice that is potentially useful and unlikely to lead completely astray. Even those who have extreme disagreements about Lewis's conclusions often allow that he produced high-quality philosophy. So Lewis is one place to start for those looking for an exemplar of contemporary philosophy. My view is that we can do even better, methodologically speaking, than Lewis, but even if we could do as well, that would be doing well indeed.[5]

Notes

1 Lewis did defend a causal theory of explanation: but cases like the collapsing star meant that his theory was in terms of causal information more generally, not just the citing of causes.
2 Lewis discusses this method explicitly in Lewis 1983, 78–95 and 1999, 248–61. I discuss this aspect of Lewis's method not only in Nolan 2005 (213–27), but in Nolan 2009.
3 I make this point in Nolan 2009 (280–2).
4 There are a number of Bayesian "convergence theorems" around, but even they do not prove that *any* set of admissible priors will *definitely* converge with *all* others, even in the limit.
5 Thanks to Chris Daly and Wolfgang Schwarz for discussion of Lewis's method.

References

Braddon-Mitchell, D. & Nola, R. (Eds.) (2009). *Conceptual analysis and philosophical naturalism*. Cambridge, MA: MIT Press.

Carnap, R. (1963). Replies and expositions. In P.A. Schlipp (Ed.), *The philosophy of Rudolf Carnap*, pp. 859–1013. LaSalle, IL: Open Court.

French, S. & McKenzie, K. (2012). Thinking outside the (tool)box: towards a more productive engagement between metaphysics and philosophy of physics. *European Journal of Analytic Philosophy*, 8, 42–59.
Goodman, N. (1954). *Fact, fiction and forecast*. Cambridge, MA: Harvard University Press.
Jeffrey, R. (1983). *The Logic of Decision*, 2nd ed. Chicago: University of Chicago Press.
Lewis, D. (1983). *Philosophical papers*, Vol. I. Oxford: Oxford University Press.
Lewis, D. (1986a). *Philosophical papers*, Vol. II. Oxford: Oxford University Press.
Lewis, D. (1986b). *On the plurality of worlds*. Oxford: Blackwell.
Lewis, D. (1989). Finite counterforce. In H. Shue (Ed.), *Nuclear deterrence and moral restraint*, pp. 51–114. Cambridge: Cambridge University Press.
Lewis, D. (1991). *Parts of classes*. Oxford: Blackwell.
Lewis, D. (1995). Ern Malley's namesake. *Quadrant*, March, 14–15.
Lewis, D. (1998). *Papers in philosophical logic*. Cambridge: Cambridge University Press.
Lewis, D. (1999). *Papers in metaphysics and epistemology*. Cambridge: Cambridge University Press.
Lewis, D. (2004). How many lives has Schrödinger's cat? *Australasian Journal of Philosophy*, 82(1), 3–22.
Lewis, D. & Lewis, S. (1970). Holes. *Australasian Journal of Philosophy*, 48, 206–12. Reprinted in Lewis (1983), pp. 3–9.
Moore, G.E. (1959). Proof of an external world. In G.E. Moore, *Philosophical Papers*, pp. 127–150. London: George Allen & Unwin.
Nolan, D. (2005). *David Lewis*. Chesham: Acumen.
Nolan, D. (2009). Platitudes and metaphysics. In D. Braddon-Mitchell and R. Nola (Eds.), *Conceptual analysis and philosophical naturalism*, pp. 267–300. Cambridge, MA: MIT Press.
Smith, M. (1994). *The moral problem*. Oxford: Blackwell.
Stalnaker, R. (1984). *Inquiry*. Cambridge, MA: MIT Press.
van Fraassen, B. (1989). *Laws and symmetries*. Oxford: Clarendon Press.

5

On Metaphysical Analysis

DAVID BRADDON-MITCHELL AND KRISTIE MILLER

5.1 Introduction

Metaphysics is largely an *a priori* business, albeit a business that is sensitive to the findings of the physical sciences. But sometimes what the physical sciences tell us about our own world underdetermines what we should think about the metaphysics of how things actually are, and even how they could be.

This chapter has two aims. The first is to defend a particular conception of the methodology of *a priori* metaphysics by, in part, exemplifying that methodology and revealing its results. The second is to present a new account of holes. These two aims dovetail nicely. We are independently interested in providing a better analysis of the concept 'hole' that yields a more plausible metaphysical story about holes. But focusing on holes is also a good way to explore the methodology we endorse: for this is an area of metaphysics that is sufficiently self-contained and narrow in focus that it provides a manageable case study, while at the same time raising interesting and deep issues about the nature of space. Ultimately we defend a new, functionalist, analysis of holes, which, unlike its rivals, neither misidentifies nor renders us implausibly eliminativist about holes under various different metaphysical suppositions about the nature of space. In the process, we set out the complex relations between the intension of "hole," and its extension at various worlds under different suppositions about the nature of space. In explicating these relations our account exemplifies what we take to be the core methodology in *a priori* metaphysics.

What is this methodology? Metaphysics is, we think, in large part a matter of constructing *conceptual maps*. These maps that tell us what we should say about the referent of some term – in this case "hole" – given different suppositions about the referent of some other term – in this case "space." Those who are sympathetic to conceptual analysis, or to something like the revival of naturalized conceptual analysis following the publication of Jackson's *From Metaphysics to Ethics* (1998) (sometimes called the Canberra Plan) will be familiar with the idea that determining the intension of a term, or the right analysis of a concept, involves equilibrating over what we are disposed to say

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

about the referent of the term under different suppositions about which worlds are actual and counterfactual.

But building a conceptual map involves more than just armchair reflection on our semantic dispositions. We want to know what we would say about the referent of, in this case, 'hole,' under different suppositions about the nature of space. Some of these suppositions will be ones according to which space is one way, and necessarily so, and other suppositions will be ones according to which space is a different way, and necessarily so, and others will be ones according to which space is one way in some worlds, and another way in other worlds. So many of these suppositions do not reflect any possible way space could be. What the map shows us is the structure of conceptual, not logical, space. Some proper part of the map represents logical space, but often the map does not tell us which part that is.

Conceptual maps are important precisely because often we do not know which part of the map represents logical space, because often we do not know whether a metaphysical view is logically necessary, impossible, or contingent. Since we know that some part of the map represents logical space, we know that *if* it turns out that we are located just *here* on the map, then *these* are the metaphysical claims to which we will be committed. Sorting out these various conditional claims – if *this* bit of the map represents the logically possible worlds, then *here* is how the metaphysics goes – is one of the central roles of *a priori* metaphysics. Of course, metaphysicians can then go on to debate about which part of the map represents the possible worlds, but, arguably, undertaken *a priori* in some cases that investigation is more fraught. In addition, conceptual maps capture *something* important about our concepts – whether one wishes to call it content or not – that possible world semantics leaves out. For there seems to be something perfectly sensible in asking what we should think about the nature of holes if, necessarily, space is substantival, even if we suppose that "space is substantival" expresses a counter-possible. Indeed, we can all more or less agree about what the right account of holes would be if, as we think *per impossible*, space were that way. This is not to say that talk of such maps commits one to an ontology of impossible worlds, or that it commits one to any sort of inflationary account of hyperintensionality, according to which there is a difference in content between necessarily true propositions, and also, on some stronger accounts, between necessarily false propositions. These would be further commitments. We do not defend any thesis about impossible worlds or propositional content. We suggest only that metaphysics, in its *a priori* mode, is essentially involved in creating and exploring these conceptual maps.

How does this idea mesh with the methodological prescriptions due to David Lewis, which have so influenced philosophers (Lewis 1972)? Lewis's view can be taken as a kind of epistemic conservatism. Start with what we actually believe – the so-called platitudes[1] together with best science (which might be thought of as what 'we' believe), and then sort out a coherent theory which preserves as much of this as possible. Our take is nuanced differently in ways that correspond to these two components. There is a gradation of ways in which we might be concerned with coherence. At one end it is purely syntactic. There's only a contradiction in our views if there is an expression and its formal negation in the deductive closure of the sentences that express our platitudes. At the other it is purely semantic: we take as fixed the contents of all these beliefs, and they require revision if the content of them is inconsistent. But there is plenty in between. One might take the content of the platitudes to determine whether our beliefs are consistent or not, but take it that many of the belief-reporting sentences that we hold true have unsettled content. We may settle on the content of those sentences (or the psychological analogues of sentences) in part in response to various meta-attitudes, including preferring the sentence to be analyzed so as to be true, or as to be false, or to be inconsistent with some other sentences. On our view part of philosophical and metaphysical analysis is spelling out the constraints on how to assign the very content that will determine whether our beliefs are consistent, and the circumstances under

which we would justifiably assign one content, and the circumstances under which we would assign another.

The other half of this is the role played by science, and by non-final metaphysics. Metaphysics is at least sometimes influenced by science, but it's helpful to know how it would be influenced under different assumptions about what is the case given the fallibility of the sciences. So our maps have a kind of conditional structure. That conditionality also extends to metaphysical claims in different domains. If you are interested in one topic it might well end up, as in the topic of the current paper, to depend on another. What we think about holes may depend on what we think about the metaphysics of space. But we will get nowhere if we have to wait for the one true metaphysics of space. Many accounts of space are empirically equivalent, and even if there are reasons for preferring one they are unlikely to be apodictic. And, for all we are justified in asserting, it might be that all the false views are impossible or even contentlessly incoherent. So we have to be able to say something about what we should believe conditional on many different accounts, some of which may be strictly contentless, or at least impossible. In which case we need an account of the normative connections between our beliefs which is not really about either what is formally inconsistent, or even more broadly inconsistent in terms of content, but rather in terms of norms of reasoning behind a hyperintensional veil, where we are unsure not only what the content of all of our thoughts and platitudes is, but even whether some of them have intentional content or are even coherent. We'd like to take all this to be an extension of Lewis's view, but perhaps that's partly the common philosopher's desire to have their view turn out to be a lineal descendent of their philosophical hero's.

We do not pretend that our discussion of different views about the nature of space is exhaustive, for our aim is not the ambitious one of creating a complete conceptual map for the metaphysics of holes. Rather, we hope to both motivate our claim that maps are not only useful, but core components of metaphysics, whilst at the same time giving a flavor for the map in question, which ultimately leads us to a new analysis of "hole."

5.2 Some Background on Holes

There are three different kinds of accounts of the nature of holes. The first, deflationary, sort of accounts, identify holes with hole-linings, or more generally with surfaces or parts of paradigm material objects. The Lewises' (Lewis and Lewis 1970) account is the best exemplar of this sort. The Lewises identify holes with what we would normally think of as "hole-linings," that is, with a certain part of the surface of the material object in which we would tend to say there is a hole and which we would tend to say "lines" the hole.

The advantage of the Lewisian account is that we identify holes with perfectly respectable ontological entities: surfaces of material objects. Indeed, the Lewisian account can be seen as motivated by the desire to avoid identifying holes with mysterious absences. The account succeeds in rendering holes perfectly ontologically respectable, however, at the cost of failing to do justice to our intuitions about the location and dimension of holes.

It was in the shadow of this failure that Casati and Varzi (1994) developed their account of holes, an account that was sensitive to intuitions about holes. On their account, holes turn out to be located roughly where we think they are. The trade-off for preserving this folk conception was that they held that holes were a fundamentally different kind of entity to material objects. On their view, holes are a particular kind of *immaterial being*: immaterial beings with particular topological and relational properties. Call this the *immaterial beings view*. To make sense of this notion, Varzi and Casati embrace

the idea that we should be realists about *space*. On their view space can be qualified – it can have certain intrinsic properties – or it can be bare or unqualified – lacking such intrinsic properties. Then, they claim, material objects are composed of qualified space. Immaterial objects are composed of unqualified space. Once we add unqualified space into our ontology, we have something ontologically respectable with which to identify holes. Moreover, since even unqualified space has dispositional properties – such as being able to be qualified – we can explain how immaterial beings, and holes specifically, can have certain dispositional properties.

The third account, due to one of the present authors (Miller 2007) recommends that we identify holes with *material beings* of a certain non-paradigmatic sort. Call this the *material beings view*. On this view, holes exist where Varzi and Casati and the folk think they do, but they are not *immaterial* beings. Unlike Varzi and Casati's account, which *identifies* material beings with regions of qualified space, the material beings view remains silent about the underlying nature of material beings. It says only that *O* is a material being iff *O* is composed of matter. This leaves open the nature of the composition relation, and also leaves it open that matter might have a different nature in different worlds, for instance, that in some but not other worlds it might be qualified substantival space.

This view is motivated by what physicists tell us about our world. Namely, they tell us that there are no regions of space across which the fields have a uniform value of zero: there are energy fluctuations across all such regions. So there are no regions of unqualified space, in Varzi and Casati's sense. Or, to put it more simply, there are no *empty* regions of space. But then if Varzi and Casati are right that holes are composed of unqualified space, it turns out that actually there are no holes. Since presumably we do not want to be hole eliminativists, we have reason to think that holes are not immaterial beings. We also have reason to think that holes are not hole-linings. If a region is occupied just in case that region is composed of qualified space, then every region of our world is occupied. Then plausibly we can hold that the region where we are tempted to say there exists a hole is occupied by a perfectly respectable, albeit non-paradigmatic, material being.[2] And that material being is a better candidate with which to identify the hole than is a hole-lining. Hence we should conclude, according to Miller, that holes are non-paradigmatic material beings.

5.2.1 The Next Move

At this point in the debate Varzi and Casati had a response to the claim that holes should be identified with non-paradigmatic material beings. Even if, they said, in our world space is always qualified, this merely shows that it is *contingent* that in our world every hole is occupied.[3] It does not mean that we should identify holes with occupants, and hence with material beings. After all, our world *could* have been such that there are holes that are not occupied by *anything*.[4] So holes ought not to be identified with what merely contingently occupies them.

Put very roughly, the debate between the three views seems to hang first on whether we think that worlds are composed of space and, if so, whether that space is such that if it is qualified, it makes sense to suppose that we could *remove* its qualities and be left with unqualified space and hence with immaterial beings with which to identify holes. To resolve these issues, we need to consider some accounts of the nature of space. In what follows we tease out the details of some such accounts by starting with a well-worn distinction between two broad conceptions of the nature of space: *substantivalism* and *relationism*. In the following section we consider what implications each of these accounts has for an account of holes. We will not argue for any particular account of space. Rather, to develop a conceptual map we intend to explore how the different analyses of 'hole' fare under different suppositions about the nature of space.

5.3 The Nature of Space

Broadly construed, substantivalism is the view that space is a real *thing* or *substance*. Substantivalism is typically contrasted with relationism, the view that space is *not* a real thing or substance, but rather, that there exist only concrete objects and events[5] and primitive spatial and temporal relations that hold between them. According to relationists, spatial relations hold directly between objects, they do not hold, as the substantivalist suggests, in virtue of objects being related to a third thing – space.

Relationism comes in two general versions, an *actualist* and a *modal* version. *Actualist relationism* holds that there exist only actual objects and the actual spatial relations that hold between them.[6] *Modal relationism* holds that there exist actual objects, and the actual and possible spatial relations that hold between them.[7] That is, modal relationism appeals to *possible* spatial relations, relations that *would* hold *were* there certain objects and events at certain locations.

For the actualist, literally nothing – including locations – exists *between* objects: there is merely the holding of non-zero distance relations between those objects. To talk of a location at which no object exists is, at best, to claim that had there existed an object with certain spatiotemporal relations to other objects, there would have existed such a location. We can define actualist relationism as follows:

ACT REL: Actualist Relationism is true at a world *w* iff
(a) all spatial properties and relations in *w* are properties of, or relations between, objects in *w*.
Some take the following to be entailed by (a). If it isn't, then it is a separate clause:
(b) there exists a location *L* in *w* iff there exists some object at *L*.

The modal relationist, on the other hand, can accommodate the notion of unoccupied locations. In a sense, the modal relationist constructs locations out of, or perhaps better, reduces locations to, the existence of actual and possible objects.[8] It is not clear quite how this is to work. Though Butterfield (1984) rejects modal relationism, he suggests that the relationist identify unoccupied locations[9] with sets of objects or events that occupy that location in some world or other. Since there are very many possible objects or events that could have occupied a particular location, we cannot identify a location with any particular possible object. So what does it mean to say that the relevant possible objects *occupy the same location* in other possible worlds? Butterfield defines the equivalence relation as "is according to the substantivalist at the same location as." Even setting aside Butterfield's worries about this proposal – namely for one, that an object cannot exist at two different locations in distinct worlds on pain of identifying distinct locations (Butterfield 1984, 106) – it hardly seems ideal to pick out the same locations in different worlds by deferring to what the substantivalist would say about those worlds. That point seems especially pressing if one thinks that relationism is necessarily true. For then substantivalism is incoherent, and defining the relevant set of possible objects in terms of substantivalist theory would be like building a house on sand, at least if we take Butterfield to be defending a kind of realism about locations. Of course he might better be read as a kind of fictionalist about the same location relation, in which case it is less problematic if the substantivalist fiction is incoherent. We do not take it as our task to adjudicate this matter here, so we will simply assume that some sense can be made of modal relationism. We will define modal relationism as follows:

MOD REL: Modal Relationism is true at a world *w* iff:
(a) all spatial properties and relations in *w* are properties of, or relations between, objects in *w* or between objects in *w* and objects that are metaphysically[10] possible at *w*.
Some take the following to be entailed by (a). If not it should be a separate clause:

(b) there exists a location *L* in *w* iff there exists some object at *L* or it is metaphysically possible at *w* that there is an object at *L*.

In contrast to the relationist, who only posits relations between objects, the substantivalist posits relations between objects and space. Substantivalism is often expressed as the view that space has certain intrinsic properties as do the objects located within space, and the intrinsic properties of the former are independent of the intrinsic properties of the latter. Space exists independently of the objects that are located within it, and it is in virtue of being related to a particular region of space that an object is located where it is.

Unsurprisingly, metaphysical accounts of space, both actual and counterfactual, have been structured by features of different accounts of the nature of actual spacetime as revealed to us by the physical sciences. Theories about the geometry of our world, from Newtonian through neo-Newtonian to general relativistic, begin by positing a set of points and then adding in further structure. They add in a topological structure, which yields a *manifold*, and then they add in geometric structure, where the geometric structure is given by a *metric* that specifies the distances between the points.

Our current best science appeals to the theory of general relativity. Models of that theory (which represent physically possible worlds for that theory) are defined in terms of a triple of M, g and T. M represents the manifold, in this case a four-dimensional manifold of continuum many points. The metric, represented by g, defines the metric relations among the points, and in this case includes the gravitational field, and T, the stress–energy tensor, represents the distribution of matter and energy. Rival versions of substantivalism emerge when substantivalists identify actual spacetime with different combinations of M, g and T. Once this identification is in place, it is generally assumed that the underlying *form*, though not the specific features, of all possible spaces will mirror that of actual space.

So, for instance, some, like Earman and Norton (1987), identify actual substantival spacetime (if there is any) with M, a view often known as *manifold substantivalism*. Actual spacetime is identified with a particular topological arrangement of points – a four-dimensional manifold – and from there it is plausible to suppose that any possible substantival space will be identified with a manifold, though in other worlds that manifold might be composed of numerically distinct points from the actual points, and might have a different topology to the one that actual spacetime has. We can define manifold substantivalism as follows:

MS: Manifold substantivalism is true at a world *w* iff (a) there is substantival space in *w* and (ii) that space is identical to a collection of continuous points with a particular topology – a manifold *M*.[11]

Another way to think of manifold substantivalism is as a kind of topological essentialism. We identify the space in some world with a collection of points, and we ask ourselves under what conditions we would have numerically the same space. The answer is that it is essential to those points that they have the topological relations that they do, that is, it is essential to being the same point in some world, that that point exist in topologically the same manifold. Hence a different topology yields a numerically different space.

Other substantivalists identify actual spacetime with M plus g: the four-dimensional manifold plus the GTR metric. Then they identify substantival space more generally with a particular collection of points with a particular topology and metric. Defended by Hoefer (1996) and Maudlin (1988), this view is sometimes known as *metric field substantivalism*. The most common version of metric field substantivalism does more than simply identify space with the manifold and the metric, since such an identification would allow that we can have the very same space even if we distribute the points

on the manifold differently, but in a way that preserves the metric.[12] The more common version holds that it is an essential property of each point that it has the relational properties it does, and hence it is an essential property of space that each point is located where it is in relation to other points.[13] By metric field substantivalism we will mean this stronger view that:

> **MFS:** Metric field substantivalism is true in any world *w* iff (a) there is substantival space in *w* and (b) that space is identical to a collection of continuous points with a particular topology *M* and metric *g*, such that for each point the relational properties of that point conferred by the metric are essential properties of that point.

Another way to think about metric field substantivalism is as a kind of metric essentialism. We identify the space in some world with a collection of points, and ask ourselves under what conditions we have numerically the same space. The answer is that it is essential to those points that they have the topological and metric relations that they do. Not only is it essential to any point that it exist in topologically and metrically the same manifold, but also that it have the particular metric relations to other points that it has – so even a manifold with the same topology and metric but with the points moved around would be numerically different space.

The final version of substantivalism that we will consider is *tensor field substantivalism*. This is the view that actual spacetime should be identified with M, g and T: with a set of continuous points on a manifold with a particular topology and metric, and a particular distribution of matter and field properties. More generally:

> **TFS:** Tensor field substantivalism is true in any world *w* iff (a) there is substantival space in *w* and (b) that space is identical to a collection of continuous points with a particular topology, *M*, and metric *g*, such that each of those points has all of its properties essentially.

According to tensor field substantivalism, there is *in some sense* no distinction between space, the container, and the objects that exist within that container, insofar as there is nothing more to either the objects or the space than the collection of points and their properties.[14] If tensor field substantivalism is true, then there are no qualitatively distinct worlds at which the same space exists: for to change the distribution of matter in any way is for there to be numerically distinct space.

The question now arises as to what modality one should attach to these views. One might be a *necessitarian*, holding that whichever account of the nature of space holds, it holds of necessity. Or one might be a *contingentist*, holding that at some worlds one account is true, and at other worlds other accounts are true. There are different versions of contingentism. Not all contingentists might think that *every* account of the nature of space we've just described reflects a possible way space could be. What we'll call "strong" contingentism accepts just that.

Our methodology is to evaluate each of the analyses in the light of each of the different suppositions about the nature of space that we have so far outlined. In a sense then, we are evaluating the analyses *as if* strong contingentism is true. This is not because we think it the most plausible view. Indeed, we think it much more likely that if substantivalism is true at any worlds, then the very same *kind* of substantivalism – manifold, metric field, or tensor field – is true at all those worlds. But we are interested in evaluating different *analyses* of the concept 'hole'. If an analysis of 'hole' is sensitive to the right account of space, then we want to evaluate it under the broadest range of suppositions about space.

One might, however, object. If space is one way and necessarily so, then as long as space is a way that makes one's preferred analysis correctly identify holes across the possible worlds, then the fact that some *other* account of space would have resulted in one's account misidentifying or

eliminating holes can be of no significance. That one's analysis fails in certain *impossible* worlds is inconsequential.

But the metaphysics of different accounts of space is controversial. Even if one is a necessitarian about one's preferred account, one should surely give reasonable credence to being mistaken about either the modality or the details of the correct account. Yet once one gives credence to alternative accounts being true, one should be worried if one's account of what it is to *be* a hole would eliminate or misidentify holes if an alternative account turned out to be true.

It is important to pull apart the *analysis* of our concept 'hole', that is, the analysis that yields the intension of "hole," from the extension of that term at different worlds given one's views about the nature of space. The former is the result of *a priori* analysis, the latter give us the metaphysics of holes when combined with that analysis.

Given that one believes that a particular account of space is necessarily false, one should expect one's views about the *extension* of holes to be false in worlds where that account turns out, *pace* your expectations, to be true. But one should not expect that in such worlds one's *analysis* of what it is to be a hole will result in the misidentification or elimination of holes in that world. For that, surely, is to say that one has the wrong analysis.

In what follows we consider how each of the analyses fares given different suppositions about the nature of space. For simplicity, we express this in terms of the traditional locution of considering some world *w* at which that supposition is true. Since strong contingentists think all of these suppositions are ways a world could be, they will read this locution in the usual manner, as considering some *possible* world *w*. Necessitarians, however, should read the locution in terms of considering some *epistemically possible* world *w*. For necessitarians are considering what they should say about holes in some world if it turned out, contrary to the necessitarian expectation, that some other account of the nature of space is true (contingently or necessarily) and hence the world in question is possible.

By considering all of these suppositions we begin to create a conceptual map that tells us important things about our concept 'hole', and furnishes us with all that can be said, *a priori*, about holes.

5.4 Space, Worlds, and Holes

Let us recall briefly the three analyses of holes that are on offer. The first, Lewisian, account is that holes are material hole-linings. The second, immaterial beings, account, due to Varzi and Casati, is that holes are immaterial beings with a certain topology and relational properties. The third, material beings, account, due to Miller, is that holes are non-paradigmatic material beings with a certain topology and relational properties.

As we will see, the problem with all three analyses is that for each there are worlds at which that account seems wrongly to locate the holes, or worse, to implausibly render hole eliminativism true at that world.

5.4.1 Relationism and Holes

Let us begin by considering relationist accounts of space. Suppose actualist relationism is true in w_1. Further, suppose that w_1 is what we will call a *non-plenum world*, where, intuitively, *plenum* worlds are worlds in which every region is occupied by some sort of matter or energy field, and *non-plenum* worlds are worlds in which there are unoccupied regions. Substantivalists can straightforwardly define plenum and non-plenum worlds in terms of the occupation of regions. But in order to be

neutral between whether a world is one in which relationism or substantivalism is true, we define non-plenum worlds as follows:

> **Non-plenum:** *w* is a non-plenum world iff there exists an *unsurrounded object* where an unsurrounded object *O* is one such that there is a point on its border, *P* and a direction *D*, such that the closest object not identical to *O* in that direction is a non-zero distance away.

Intuitively, in non-plenum worlds the location at which we would pre-theoretically say there is a hole might be wholly occupied by matter, or it might be partially occupied by matter, or it might be entirely unoccupied by matter. We will primarily be interested in *occupied* and *unoccupied holes*. Occupied holes are ones that, according to our pre-theoretic judgements as to their location, are completely occupied by matter. Unoccupied holes are ones that, according to our pre-theoretic judgements as to their location, are completely unoccupied by matter. Talk of unoccupied holes is straightforward for the substantivalist, but not for the relationist. Below we give a feel for how a definition consistent with relationism might proceed, before returning to consider this definition and the idea of an unoccupied hole in the context of the ensuing discussion in terms of actualist relationism.

> **Unoccupied hole:** an unoccupied hole *H* in an object *O* is one such that on every point *P* on the surface of *O* which is on the boundary of *H*, in every direction from *P* the nearest object (if there is one) in that direction that is within the hole opening is a part of *O*.
> **Hole opening** = def that part of what, intuitively, we would describe as the hole boundary (if any), which is not part of the object boundary.

Unlike plenum worlds, non-plenum worlds offer the possibility of unoccupied holes. So, let us consider w_1, which, recall, is a non-plenum actualist relationist world. Consider what seems to be an unoccupied hole, H_1 in object O_1 in w_1. That hole cannot be either an immaterial being or a non-paradigmatic material being. There is no space in w_1, and hence there is no immaterial being in the relevant location. Nor, by definition, is there any matter in the location where we think the hole exists, and hence there is no material being at the location. So, in actualist relationist worlds, both the immaterial and the material beings views render us eliminativists about holes like H_1. In w_1, the right account of holes seems to be the Lewises', which identifies the hole "in" O_1 with a part of O_1's surface – the hole-lining.

This makes talk of an "unoccupied hole" a bit misleading, since it is not that there is a hole with certain dimensions which is unoccupied by matter: the hole turns out just to be the surface of the object that has the hole. But that is what we should expect. If actualist relationism is true, then even *defining* an "unoccupied hole" is problematic. We appealed to the idea of a hole opening. But that requires that we are able to talk of the parts of a hole boundary that are not parts of the object boundary, and to do that it looks like we need to be able to talk about unoccupied points – the points that are on the hole opening, but where no matter is located. And that is precisely what actualist relationism will not allow us to do. For there are no such points, and no such locations. Thus we cannot define a hole-opening and hence cannot define an unoccupied hole. That is not surprising, we cannot even talk about *the location* at which we are tempted to say there exists a hole, because there is no such location. Strictly speaking we should say that, in worlds like w_1, there are no unoccupied holes, though there are holes in unsurrounded objects on the parts of the surface of those objects that do not border any other objects. However, for ease of exposition we will continue to talk of unoccupied holes in actualist relationist non-plenum worlds. The point is that, on pain of eliminativism, such holes must be identified with surfaces of material objects just as the Lewises' suggest.

Now consider w_1*, which is like w_1 in being a world where actualist relationism is true, except that w_1* is a *plenum world*:

Plenum: *w* is a plenum world iff there exist no *unsurrounded objects* in *w*.

Plausibly, ours is a plenum world. It is a world where, for instance, holes are filled with gas, or water or energy fluctuations, or some-such. In w_1* if there are any holes, they are *occupied holes*, where *H* is an occupied hole iff there exists matter in exactly the location where, intuitively, we would say there exists a hole. (Of course, if one is an actualist relationist then there is only such a location if there is matter there.)

In w_1* both the Lewisian analysis and the immaterial beings analysis are unsatisfactory. The former identifies the holes in w_1* with parts of the surfaces of the paradigm objects that "house" the hole. But that seems wrong, since there is a better candidate in w_1* to be identified as the hole, namely the entity that occupies just the location and has just the dimensions that we would attribute to the hole: a non-paradigm material being (or object). The immaterial beings analysis offers an even worse outcome: it renders us eliminativist about holes in w_1*, since there are no immaterial beings in w_1*. Indeed, since for Varzi and Casati material beings are composed of qualified space, their view has the much worse implication that there are no material beings in w_1*, or indeed, in any relationist world. The virtue of the material beings view is that it does not identify material beings with regions of qualified space. It can still maintain that there is a material being in the region in question – just not a material being composed of qualified space. Thus in worlds like w_1* – actualist relationist plenum worlds – the material beings analysis correctly identifies the holes where the other two analyses fail.

Consider now modal relationism. As we see it, whether a world is an actualist relationist world or a modal relationist world will have no effect on which analysis successfully identifies the holes at that world. This is clearly true in plenum worlds. Since all possible relations are actual relations, the two views collapse into one another in plenum worlds. In non-plenum worlds matters are less clear. Consider a world w_2 that has the same distribution of matter as w_1 (and is thus a non-plenum world), but where modal relationism is true. In w_2 we can sensibly talk about the unoccupied region surrounding O_2 (the counterpart of O_1 in w_1). So it is much more straightforward to talk about an unoccupied hole in O_2. That might lead us to suspect that a non-Lewisian analysis is feasible in such a world.

The difficulty is that there is nothing in the unoccupied region with which to identify the hole: either material or immaterial. One option would be to identify the hole with a true void: an unoccupied region in a world without substantival space. We are prepared to follow the modal relationist in holding that we can *talk* about unoccupied regions without being committed to the existence of substantival space, and hence that, in some *minimal* sense, perhaps, there exist true voids. But the sense in which such voids exist is not one in which there exists some substantial nothingness with which one can identify a substantial absence – namely a hole. Our claim is not that if there were true voids, then we should identify holes with those voids, but there are no voids (and necessarily so) thus we cannot so identify. For that would be to adjudicate the question of whether there are possible worlds with voids, a task we explicitly eschew. Rather, we think that even if modal relationism is true at some worlds, those are not worlds with true voids in the sense that would be required for there to exist something with which to identify a hole. But should the reader disagree, we leave it as an exercise for him/her to amend our conditional analyses in the light of the view that in modal relationist worlds, holes are to be identified with true voids. This will make the analyses more complex, but nothing that we say otherwise will be changed. On the assumption that we are correct, however, we will simply talk about "relationism," where this is understood to mean either actualist or modal relationism.

Then the Lewisian analysis is vindicated in any relationist world with unoccupied holes, and the material beings view is vindicated in any relationist world with occupied holes.

5.4.2 Substantivalism and Holes

Varzi and Casati are clear that their view presupposes that substantivalism is true. Their view is that we should identify holes with immaterial beings that are regions of unqualified space. But what, in the light of the different sorts of substantivalism that we have outlined, does that amount to and how does their account fare in comparison with the material beings view? We begin by considering tensor field substantivalism. This is probably the least common kind of substantivalism, but it is worth considering because of the light it sheds on the debate between the immaterial beings and the material beings analysis of 'hole'. First though, we should note that most of the discussion of different kinds of substantivalism focuses on the different implications for the material beings and the immaterial beings analyses. For obvious reasons. We think it pretty plain that if a world has substantival space of some kind or other, then both the material and the immaterial beings analyses are preferable to the Lewisian analysis. If there is some entity in the region in which we tend to say there is a hole, then that entity is a better deserver of "hole" than is a hole-lining. Consideration of different kinds of substantivalism is thus primarily directed at discerning how the other two analyses fare under different suppositions about the nature of substantival space.

5.4.2.1 Tensor field substantivalism Suppose tensor field substantivalism is true in w_3, and that w_3 is a plenum world. Suppose we identify the holes in w_3 with regions of substantival space as Varzi and Casati demand. Since all regions of space in w_3 are decidedly *qualified*, in Varzi and Casati's sense, have we identified the holes with material or immaterial beings?

In the light of the possibility of plenum worlds, Varzi and Casati shift their view of immaterial being from the claim that immaterial beings are regions of unqualified substantival space to the view that immaterial beings are regions of substantival space, such that if the region in question is qualified, then it is possible to remove the qualities. More carefully, the view seems to be that x is an *immaterial being* iff x is identical to a region of substantival space, and if that region of space has intrinsic qualities, then we could remove all of the qualities from that region of space and x would still exist in that region. Holes are immaterial beings, they maintain, because in a plenum world we could remove all of the matter and still have the same hole in the same region. Their newer understanding of immaterial beings then suggests they should say that x is a *material being* iff x is identical to a region of qualified space, such that we could not remove all of the matter from that region of space and x would still exist in that region.

Recall that tensor field substantivalism is the view that each spacetime point has all of its properties essentially. This entails that it is essential to space in w_3 that it have the matter field it does and hence entails that there is no world that has the same space as w_3, but where the matter has been removed from the regions occupied by the holes in w_3. It is not, therefore, merely contingent that the holes in w_3 are occupied. Rather, given the definitions just considered, it turns out that all regions of tensor field substantival space are identical to material, not immaterial beings. There are no immaterial beings in w_3. Hence the analysis makes us eliminativists about holes in w_3.

There are some options open here. We could think of the immaterial beings view as simply telling us that holes should be identified with regions of substantival space. It could then be a *further* posit of the view that these regions will turn out to be immaterial beings, on the assumption that substantival space is a certain way, namely a way that renders it possible to have numerically the same space unoccupied by matter. That further posit could turn out to be false, without making us eliminativists about holes. It would simply yield the result that we identify holes with a *material* rather than an

immaterial being. That is not, we think, what Varzi and Casati intended, but considering the view will help clarify some important issues that will arise later.

Suppose we do identify the holes in w_3 with regions of substantival space, setting aside the problem that such regions are *not* immaterial beings. Then there are other problems. Perfectly ordinary counterfactual conditionals about holes come out false. Claims such as "if I had dug deeper, this hole would have been larger" are false, since the closest world where I dig deeper has a different matter field from w_3, and hence is a world with numerically distinct space and hence numerically distinct holes.

What then, of the material beings view? The material beings view will certainly not want to *identify* material beings with regions of qualified space. After all, this would render it eliminativist about material beings in any worlds that lack substantival space. She might want to say that in worlds with substantival space, material beings are *constituted* by, or *contingently identical* to, regions of qualified substantival space for they are composed of matter, and in such worlds matter is constituted by, or contingently identical to, regions of qualified space. Then in worlds without substantival space the view might hold that material beings are composed of matter by being constituted by, or contingently identical to, a mereological fusion of matter particles, or to the values across the matter field. For simplicity and parity with Varzi and Casati's view, however, we will assume that, in worlds with substantival space, material beings are composed of regions of qualified substantival space. In particular, we assume that in worlds with substantival space, material beings are contingently identical to regions of qualified space, though everything we say can be translated, *mutatis mutandis*, into talk about constitution. Further, for clarity we will use the phrases "straightforward identity" or "straightforwardly identify" to refer to the standard, non-contingent sort of identity. More specifically, we will say that x is a material being just in case:

> *X* is a material being iff either (i) x exists in a world where there is substantival space and x is contingently identical to a region of qualified substantival space such that we could not remove all of the qualities from that region and x would still exist in that region, or (ii) x exists in a world where relationism is true and x is contingently identical to a fusion of matter particles or to a region of a matter field.

Then the idea is that terms like "Fred" that refer to an object and "Molly" that refer to a particular region of space, pick out the very same entity in one world – w_3 – but pick out distinct entities in other worlds, because Fred and Molly have different counterparts in other possible worlds. There are object-counterparts of Fred's that are not spacetime region counterparts of Molly's, and vice versa. Hence material beings can be contingently identical to different regions of space in different worlds.

On this view of material beings, the material beings account and the "amended" immaterial beings account both identify holes in w_3 with material beings, but they differ in what they will say about the counterfactuals regarding those holes. For the material beings view can allow that the very same hole exists in qualitatively distinct worlds, despite the fact that the same region of space does not. Counterfactuals such as "had I dug deeper, the hole would have been larger" will come out true: in the closest world where I dig deeper the material being that is the hole exists and is larger. It is just that that material being is contingently identical to a numerically distinct region of space from the one with which it is contingently identical in w_3.

So in plenum worlds in which tensor field substantivalism is true, the material beings analysis is clearly superior to any of the alternatives. It is superior to the Lewisian analysis, since it locates the holes in a more plausible location, giving them genuine dimensions and various other dispositional properties. It is also superior to the straight immaterial beings view, since that view results in hole-elimination. And it is superior to the amended immaterial beings view, which not only, ironically,

turns out to identify holes with material beings, but which renders false ordinary counterfactuals about holes.

On the other hand, the material beings view fares badly in non-plenum worlds in which tensor field substantivalism is true. For in such worlds there may be unoccupied holes, and hence holes that cannot be identified with any material being. In such worlds the material being analysis will render us eliminativists about unoccupied holes. In non-plenum worlds the immaterial beings view allows us correctly to identify the holes, and to identify them as immaterial beings. But so long as it straightforwardly identifies the holes with regions of, in this case, unqualified substantival space, it still faces the problem that perfectly ordinary counterfactuals about those holes will come out false.

5.4.2.2 Counterfactuals and the error theory By now we should be getting a feel for the general form of the problem facing our three analyses. We begin by supposing, for each different account of the nature of space, that it is *at least epistemically possible* that there is a world is one in which that account is true. Then we discover, for each analysis of "hole," that that analysis misidentifies or eliminates the holes in some (at least) epistemically possible world. That gives us reason to suppose that the analyses are mistaken. In general, this is the problem that the analyses implausibly leave us open to the error theory being true of holes should certain suppositions about space turn out to be possible. Call this *the error theory problem*.

We have just seen, however, that even abstracting away from the error theory problem leaves some problems for our analyses, and these problems are generated without any appeal to what some might maintain are strictly counter-possibles. This is a set of problems that arise when we try to evaluate counterfactual conditionals about holes: call it *the counterfactual conditional problem*. These two broad problems appear to be quite distinct, but furnishing a solution to the latter will help with the former. Let us see how.

We suggested that the right account of material beings would not straightforwardly identify them with qualified regions of substantival space. Rather, material beings should be thought of as contingently identical to such regions.

This suggests we might adopt a similar treatment for holes. So in a world *w* in which tensor field substantivalism is true, any hole is contingently identical to a region of substantival space, namely the space that surrounds the hole-lining and has certain dimensions and topological properties. Let us call the totality of properties relevant to determining the location of a hole the *H-properties*. These properties include the properties of the hole-lining, the topology of the hole, and other relational properties about the location of objects in the world. The idea is that we track the counterparts of any hole in *w* in terms of these H-properties. In the relevant counterfactual worlds, we locate the counterpart of the object in *w* that "houses" the hole, and in doing so we locate the hole-lining in that world. From there we can determine the dimensions and topology of the hole, and hence determine *which region of space in that world* the hole is contingently identical with. Once we appeal to an account such as this the problem of counterfactual conditionals vanishes, since holes have counterparts in the relevantly closest worlds even though those worlds have numerically distinct space. In general, we suggest the following:

> **Analysis 1:** In any world *w* in which there is tensor field substantival space, a hole *H* is contingently identical to the region of substantival space that has the H-properties.

We could think of this as a functionalist analysis of 'hole'. The functional role is captured by the H-properties. What *realizes* the H-properties varies between worlds depending on which region of tensor field substantival space is playing the role. Hence we can have the same hole being realized

by different regions of space.[15] Since, in any world in which tensor field substantivalism is true, the holes are contingently identical to regions of tensor field substantival space, it follows that occupied holes are contingently identical to material beings, and unoccupied holes are contingently identical to immaterial beings. So what is *right* about both the material beings and the immaterial beings views is not the analysis they offer, but, rather, that there are worlds in which each correctly identifies what it is that realizes the hole. Some holes are material beings, and other holes are immaterial beings. But it is not part of an analysis of 'hole' that they should be straightforwardly identified with either.

Bearing in mind this simple analysis, we can now consider some other versions of substantivalism, beginning with metric field substantivalism.

5.4.2.3 Metric field substantivalism Suppose that w_4 is a plenum world in which metric field substantivalism is true, and that it is a world in which general relativity holds. Now we follow Varzi and Casati in *straightforwardly* identifying holes in w_4 with regions of metric field substantival space. Given metric field substantivalism, there is a world, w_4*, that has the same space(time) as w_4 (in virtue of having the same metric and points) and which has *almost* the same distribution of matter except that the region at which there is a hole in w_4 is, in w_4*, unoccupied. w_4* is a non-plenum world that has an unoccupied hole at the same region where there is an occupied hole in w_4. Since w_4* has the same space as w_4, the region of space where the hole exists in w_4 *could have been unoccupied*. So it looks as though, if metric field substantivalism is true, then Varzi and Casati's account succeeds, for, even in plenum worlds, there *is* an immaterial being with which to identify the hole – namely, the region of substantival space from which we could remove all of the matter. It is just that the hole in w_4 is contingently occupied by some material being.

But immediately a problem arises: w_4* is not close to w_4. It has very different laws of nature. We know that given general relativity, g, the metric, and T, the stress–tensor field, are interdependent. For g includes the gravitational field, and hence has implications for the distribution of matter and energy. Consistent with the general theory of relativity, we cannot remove or rearrange the matter and have the same metric: different matter field, different metric. Any world that is nomologically accessible from w_4 (that is, any world that is nomologically possible if w_4 is a possible world) where there is matter absent from just the region where in w_4 there is a hole is a world with a different metric and hence a world with numerically distinct spacetime.

Thus if we straightforwardly identify a hole in w_4 with a region of substantival space, a whole range of plausible counterfactual conditionals come out false. A conditional Varzi and Casati want to endorse, such as "if I removed the matter from this region where there is a hole, there would remain the very same hole, now unoccupied" will be false. For the closest world in which to evaluate that conditional – and hence a world with the same laws of nature – is one in which the matter is rearranged, and hence the metric is changed and the spacetime is numerically distinct from spacetime in w_4. So in that world the hole does not exist. Similarly, perfectly run of the mill counterfactuals such as "if I had dug more dirt out, this hole would have been larger" also turn out to be false for the same reasons.

The material beings analysis fares no better. In plenum worlds it outperforms the immaterial beings analysis. It gets right counterfactuals such as "if I had dug more dirt out, this hole would have been larger." Since the closest world will be a plenum world (given that the w_4 is) and will have the same laws of nature, there will be a material being in that world with which we can identify the hole – a material being that in the relevant world is indeed larger. But, once again, in non-plenum worlds with unoccupied holes the material beings analysis falls prey to the error theory problem. Worse still, given metric field substantivalism, even if w_4 is a plenum world, there are *nomologically accessible* worlds with unoccupied regions (albeit worlds with numerically distinct spacetime), and hence

potentially with unoccupied holes. Yet in these quite close worlds the material beings analysis still forces us to be eliminativists about unoccupied holes.

Once again, then, even if we only focus on worlds where metric field substantivalism is true, the immaterial beings view faces the counterfactual conditional problem, while even amongst these limited worlds the material beings view faces the error theory problem.

Each could simply bite the bullet. Varzi and Casati could say so much the worse for our intuitions. If holes really are made of space, and if it turns out to be a surprising fact that our laws of nature are such that matter and energy are interdependent with the metric, then we should not be surprised to find that there are counterintuitive facts about the identity of holes across worlds. Likewise, the material beings proponent could say that we should be eliminativists about holes in worlds where there are unoccupied regions. But neither of these strategies is very attractive.

Instead, we suggest adopting the same strategy we proffered if it should turn out that tensor field substantivalism is true. Namely, if metric field substantivalism is true at a world then we contingently identify holes with regions of metric field substantival space.

> **Analysis 2:** In any world *w* in which there is metric field substantival space, a hole *H* is contingently identical to the region of substantival space that has the H-properties.

Then everything that we said previously holds here. Counterfactuals such as "if I had dug more dirt out, the hole would have been larger" come out true, since in the nearest world the very same hole exists, and is larger, it is just that in that world the hole is contingently identical to a numerically different region of space. There are two noteworthy differences between analysis 1 and 2, though, that result from the differences between tensor field substantivalism and metric field substantivalism. While analysis 1 allows that we can have the same hole, realized by different regions of space, analysis 2 *in addition* allows that we can have the same space that does not realize any hole. A token hole and a token region of space completely come apart, such that we can always have a world with one and not the other and vice versa. Nevertheless, as before, holes turn out to be contingently identical in some worlds to immaterial beings, and in other worlds to material beings. For although Varzi and Casati are right that when we identify a token hole with a region of qualified metric-field substantival space that very same space could have existed absent the occupier, on this analysis, since the identity is contingent, it turns out that the hole could have existed absent that region of space. So, contrary to Varzi and Casati, we do not identify the hole with an immaterial being under such circumstances, but with a material being, though what is true is that in some other world the hole is contingently identical to an immaterial being.

5.4.2.3 Manifold substantivalism Finally we turn to manifold substantivalism. On this view, substantival space is a collection of continuous points with a particular topology. Since topology underdetermines geometry, consistent with general relativity different worlds with the same topology can have a different distribution of matter and energy. So suppose w_5 is a plenum world in which manifold substantivalism is true, and suppose we follow Varzi and Casati in straightforwardly identifying holes with regions of manifold substantival space. Then the immaterial beings view fares rather better than on other suppositions about the nature of space. For there is a nomologically accessible world to w_5 that preserves the same topology and hence has numerically the same space as w_5, but which has a different distribution of matter. So, for any hole in w_5 that is identified with a region of qualified space, there is a nomologically accessible world with the very same space, but where that region is unoccupied by matter. It might even be that this is the closest world, so long as very minimal differences in the distribution of matter are consistent both with that region being unoccupied and

with preserving the same topology. We are not entirely convinced that such a world will be the closest, but let us give the immaterial beings view the benefit of the doubt and suppose it is. Then the counterfactual "had this region not been occupied by matter there would still have been a hole here" comes out true, and we have identified the holes in the plenum world with immaterial, not material beings. Moreover, every counterfactual such as "if I had dug deeper, this hole would have been larger" also comes out true. Likewise, the view correctly identifies the unoccupied holes in non-plenum worlds, and allows that the appropriate counterfactuals about those holes come out true.

The material beings analysis, on the other hand, misidentifies the holes. In a plenum world it identifies the holes with the material beings in that world, rather than with the space that is, contingently, occupied by those material beings. And it renders us eliminativists about any unoccupied holes in non-plenum worlds.

5.4.2.4 The dialectic Here then, is the dialectic. What holes *are* depends crucially on controversial details about the metaphysics of space. The problem for the three accounts of holes is that each builds some of this metaphysics into the very analysis of the concept 'hole'. Then, when we consider each analysis in a world where the relevant metaphysics does not hold, it turns out that we are mistaken in our identification of the holes in that world. The trick is to realize that our concept 'hole' gives us guidance to the nature of holes *given* various presuppositions about the nature of space.

Here is where we think the lessons gleaned from consideration of the nature of holes begin to take shape in a way that is relevant to metaphysics more generally. Holes, like any other phenomena that are the fodder for metaphysical theorizing, are out there in the world. We think that the right process of developing a metaphysical account of the nature of holes is essentially the *a priori* process of discovering conceptual truths about holes. It is the process of discovering what is in our concept 'hole'. But this is no straightforward matter. It is a matter of discovering what we would say about holes under starkly different suppositions about the nature of space. As we see it, the complex conceptual edifice that tells us what we should say about holes under these different suppositions exhausts all that can be said *a priori* about holes. What holes are, in the sense of what sort of entity they turn out to be, in different worlds, ultimately depends on which of these suppositions turn out to be genuinely possible and that issue might be resolved in part by further *a priori* reasoning and in part by information from the physical sciences.

5.5 Metaphysics and Analysis

With an understanding of the relationship between the metaphysics of holes and an account of the nature of space under our belts, we can now consider what to say about the correct analysis of 'hole'. In the following analyses we will use the locution "holes are identical to X." This can be read in one of two ways. It can be read in the straightforward way that implies no appeal to contingent identity and, where, if a hole is identical to X, then it is identical to X in every world in which it exists. As we noted earlier, there are serious problems that arise from embracing this conception. But there may be those who are willing to bite various bullets and declare that many of our intuitions about counterfactuals regarding holes are simply false. If so, then read in terms of straightforward identity our new analysis can be seen as a way of avoiding the more general error theory problem that arises once we consider the possibility that other accounts of the nature of space might be true. On the other hand, the locution "holes are identical to X" can be read as the claim that holes are contingently

identical to X. This latter is our preferred reading, since only then will perfectly plausible counterfactuals about holes come out true.

The following conditionals are plausible claims about holes. Let the topological and relational properties relevant to settling the location of holes be called the H-properties. Then:

1. If relationism (actualist or modal) is necessarily true, then, necessarily, holes are identical to the material beings that possess the H-properties if there are any, and to particular surfaces (hole-linings) of material beings otherwise.
2. If tensor field substantivalism is necessarily true, then holes are identical to the regions of tensor field substantival space that possess the H-properties and necessarily so.
3. If metric field substantivalism is necessarily true, then holes are identical to the regions of metric field substantival space that possess the H-properties and necessarily so.
4. If manifold substantivalism is necessarily true, then holes are identical to the regions of manifold substantival space that possess the H-properties and necessarily so.

Claim (4) is worth some discussion. We could simply have said that if manifold substantivalism is necessarily true, then, necessarily, holes are identical to immaterial beings. For if manifold substantivalism is necessarily true, then every hole is identical to an immaterial, rather than a material, being. If manifold substantivalism is necessarily true, we might think, the immaterial beings view is vindicated. In a sense that is right: all holes will turn out to be immaterial beings. But we are not convinced that this is part of the *analysis* of 'hole'. First, it is not clear why the conditional claim in (4) should have a quite different form from that in (2) and (3), given that nothing about the nature of manifold substantivalism itself or the concept 'hole' seems to suggest any such difference. Rather, we think, we make a discovery once we see that manifold substantivalism is necessarily true: we discover that, in every world, what it is that realizes the role of being a hole is an immaterial being. The immaterial beings view would turn out to be right about the metaphysics of holes, but wrong as an analysis of the concept.

Given these conditionals, there is nothing we can discover about the nature of space in a world that would make us eliminativists about holes in that world. Rather, when we discover which account of space holds, we discover what sort of entity the holes are. So the discovery that relationism is true and necessarily so is *not*, as the immaterial beings view would have it, the discovery that there are no holes. Rather, it is the discovery that, necessarily, occupied holes are material beings, and unoccupied holes are hole-linings. But that is consistent with holes being identical to regions of substantival space if it should turn out that space is substantival and necessarily so.

We can collapse the four conditionals above into a more general, user-friendly necessitarian analysis.

> **(NA)** For any kind of substantivalism *K*, if all worlds are *K* worlds then holes are identical to regions of *K*-type substantival space that possess the H-properties and necessarily so, otherwise, necessarily, holes are identical to the material beings that possess the H-properties if there are any, and to hole-linings otherwise.

It is then relatively easy to take into account the possibility (real or epistemic), that each of these accounts of the nature of space might turn out to be false – either necessarily or contingently – or turn out to be only contingently true. It seems we want to say that for any world with substantival space, *whatever kind it might be*, the holes in that world are identical to regions of that substantival space,[16] and in worlds where relationism is true, if there are any material beings that have the H-properties, then the holes are identical to those material beings, otherwise they are identical to hole-linings. This yields the more complete analysis:

(CA1) For any world *w*, if *w* is a world with substantival space of kind *K*, then the holes in *w* are identical to regions of *K*-type substantival space that possess the H-properties, otherwise holes are identical to whichever material beings possess the H-properties if there are any, and otherwise to material hole-linings.

(CA1) tells us that, in a relationist world where there are occupied holes, we conceptually prioritize material beings over hole-linings as being the referents of "hole." Holes only turn out to be hole-linings if there is no entity whatever in the relevant region: any other entity is a better candidate. If it turns out that, necessarily, relationism is true, then it is necessarily false that holes are immaterial beings. But notice that, although it would be a necessary falsehood, it would not be a conceptual truth. For the conditional nature of the concept would remain, telling us that if we were to discover that in fact there is substantival space, matters would be otherwise. Likewise, if it turns out that substantivalism of one kind or another is true at every world, then it turns out to be a necessary falsehood that holes are hole-linings, though for parallel reasons its necessity would not be a conceptual matter (even though it follows from a conceptual analysis and a metaphysical discovery). Once again it would remain epistemically possible that relationism is true (perhaps even necessarily true) and, if true, that hole eliminativism would not be vindicated.

As it stands, this analysis is not sensitive to facts about the nature of space in *our* world. It is essentially a functionalist account of holes. It tells us that holes are identical, in any world to whichever realizer has the H-properties in that world, if there is such a realizer, or else to a hole-lining. So "hole" is a non-rigid term that refers, in different worlds, to whatever possesses the H-properties, and hence to immaterial beings in some worlds and material beings in others.

But sometimes a candidate extension for a term is a significantly better deserver than any other candidates, such that we want to give that candidate some sort of conceptual priority. We might think that, if actually our world is a certain way, then holes are a certain way and necessarily so. Or we might think that, if our world is *not* a certain way, then holes are not a certain way and necessarily so. To capture this sort of conceptual priority without the threat of eliminativism we need to *conditionally rigidify*. A term is conditionally rigid if, given some number of actual candidate extensions, for each candidate extension the term would rigidify on some but not other extensions, were they to be actual. So, for instance, we might hold that immaterial beings are such good deservers of "hole," that only if actually there are no such beings should we accept that holes can be material beings or hole-linings. Thus we think that "hole" rigidly refers to immaterial beings if there are any, and to something else otherwise. If "hole" is conditionally rigid in some way, then what counts as a hole in any world is sensitive to facts about the nature of space in our world.

There are various different ways one might conditionally rigidify on ways our world could be, but two stand out as the more plausible options. The first (CA2), conceptually prioritizes as the referent of "hole" any entity that occupies the region where, pre-theoretically, we would say there is a hole. It tells us that, if actually there are such entities, then worlds in which such entities are absent are worlds where there are no holes, not worlds where holes are hole-linings. On the other hand, if it turns out that our world is one that lacks such entities, this is not the discovery that there are no holes, rather, it is the discovery that *actually holes are hole-linings*. For in that case the second clause of the conditional comes into force, telling us that, in any world, if there are any entities that possess the H-properties then they are the holes, otherwise holes are the hole-linings.

(CA2) If actually there are entities that possess the H-properties, then necessarily, any hole is identical to an entity that possesses the H-properties, otherwise holes are identical to whichever entities possess the H-properties if there are any, or otherwise to hole-linings.

The second option for conditional rigidification prioritizes substantival space as the best deserver of "hole." There are variants on this kind of analysis that rigidify on the kind of substantival space that there is actually, so that in worlds with substantival space of a different kind, there are no holes. We think this implausible. So it is consistent with (CA3) that actually there is one kind of substantival space with which holes are contingently identical, and that counterfactual holes are contingently identical with some region of substantival space or another, but not necessarily with space of the same substantival kind. But, the analysis tells us, if actually there is no substantival space then holes both actually and counterfactually are the material beings if there are any, or hole-linings if there are not.

(CA3) If actually there is substantival space, then necessarily holes are identical to regions of substantival space with the H-properties, otherwise they are whatever entities have the H-properties if there are any, or else they are the hole-linings.

In fact, we do not find these options compelling. We think that "hole" is a non-rigid term, because we think that, however our world turns out to be with respect to the nature of space, this would not give us grounds to be eliminativists about holes in worlds that are metaphysically different.

5.6 Conclusion

These series of conditionals map a segment of logical space, telling us what we will say about the actual and counterfactual extension of "hole" given that we find ourselves in different parts of that logical space. It tells us what we will say about holes given different views about the nature of spacetime. It tells us what we will say about holes given different views about whether a particular account of the nature of spacetime is necessary or contingent. And it tells us what we will say about holes given that we conceptually prioritize certain kinds of entities as being better deservers than others, or given that we don't.

Notes

1 Platitudes in this sense need not be thought of as actual conscious beliefs, or sentences taken to be true, but rather something like dispositions to form judgments or behave in different ways in different circumstances.
2 If unrestricted composition is true, then it straightforwardly follows that all occupied regions are occupied by material beings, and hence it follows that there is a material being that occupies the region where we are tempted to say there is a hole. If restricted composition is true not every occupied region is occupied by a material being, but it is still plausible that the region in question is so occupied, given that we have strong intuitions that it *is* occupied by a single entity, namely a hole.
3 We use the term "occupied" rather than filled, to distinguish the cases where a hole is occupied from cases where we would say that a hole is filled – the latter are simply cases where what it is that occupies the region is of substantially the same kind as that surrounding the hole, and hence we are inclined to say that the hole has been filled. If actual space is all qualified, then every hole is occupied, though not every hole is filled (witness holes in the road, occupied, but not filled).
4 Casati and Varzi make this point (1994, 35).
5 Henceforth for simplicity we will simply talk of objects and the relations that hold between them, but this should be read to include events.
6 Teller (1991, 364) calls actualist relationism "narrow relationism."
7 We follow Le Poidevin (2004) and Sklar (1985) in referring to this view as modal relationism. Teller (1991, 364) calls modal relationism "liberalised relationism."

8 Author (2)'s discussion of holes draws attention to the fact that there is good reason to suppose that in the actual world, every possible location is occupied – the actual world is a plenum. If that were so, then in the actual world actualist relationism and modal relationism would not come apart, insofar as they would countenance the very same locations. But given that we know that there are physically possible worlds where there is no plenum, we know that these two accounts will come apart.

9 Butterfield frames the discussion in terms of points.

10 We do not use metaphysical possibility here to indicate a special kind of possibility, but just to indicate that we are interested only in objects that exist in worlds that are like *w* insofar as (at least) modal relationism is true at them.

11 There is an issue about whether manifold substantivalists should think that spacetime is identical to a manifold, and the manifold identical to a collection of continuous points, or whether we might think that spacetime is composed of, or constituted by, a manifold. The former would be a strictly reductive account of spacetime, the latter may not be. This is an interesting issue, but for our purposes the distinction will not matter. For ease of exposition, however, we will talk only about the simpler reductive view.

12 The worry with such a view is that it is open to the hole argument, and hence fares no better in that respect than manifold substantivalism.

13 There are various ways of strengthening of weakening what is thought to be essential, for instance Healey (1995) recommends a view that has some features in common with metric essentialism but is more minimal.

14 One way to think of tensor field substantivalism is as the view that space is ontologically basic, and what it is for there to exist objects and events within space is for that space to have particular properties at particular locations.

15 Though notice that if tensor field substantivalism is true we cannot have the very same region of space failing to realize the same hole in virtue of not having the H-properties, since in order to be the same region of space it is guaranteed to have the same H-properties. So holes and regions of space only come apart in one direction: we can have the same hole realized by different space, but we cannot have the same space failing to realize the same hole.

16 So we don't think, for instance, that any of the kinds of substantival space is such an overwhelmingly better candidate with which to identify holes, that if each kind of space is possible, then only regions of *one* kind of substantival space should be identified with holes.

References

Butterfield, J. (1984). Relationism and possible worlds. *British Journal for the Philosophy of Science*, 35, 101–13.

Casati, R. & Varzi, A. (1994). *Holes and other superficialities*. Cambridge MA: MIT Press.

Earman, J. & Norton, J. (1987). What price substantivalism? The hole story. *British Journal for the Philosophy of Science*, 38, 515–25.

Healey, R. (1995). Substance, modality and space-time. *Erkenntnis*, 42, 287–316.

Hoefer, C. (1996). The metaphysics of space-time substantivalism. *Journal of Philosophy*, 93, 5–27.

Jackson, F. (1998). *From metaphysics to ethics: a defence of conceptual analysis*. Oxford: Oxford University Press.

Le Poidevin, R. (2004). Space, supervenience and substantivalism. *Analysis*, 64(3), 191–8.

Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, 50, 249–58.

Lewis, D. & Lewis S. (1970). Holes. *Australasian Journal of Philosophy*, 48, 206–12.

Maudlin, T. (1988). The essence of space-time. In A. Fine & M. Forbes (Eds.), *Philosophy of Science Association 1988*, Vol. 2, pp. 82–91. East Lansing, MD: Philosophy of Science Association.

Miller, K. (2007). Immaterial beings. *The Monist*, 90(3), 349–71.

Sklar, L. (1985). *Philosophy and spacetime physics*. Berkeley: University of California Press.

Teller, P. (1991). Substance, relations, and arguments about the nature of space-time. *Philosophical Review*, 100(3), 363–97.

6

A Lewisian History of Philosophy

ROBERT PASNAU

Great philosophers force us to rethink not only the future of philosophy, but also its past. Or, as Nietzsche more grandly puts it: "Every great human being exerts a retroactive force: for his sake all of history is put on the scale again, and a thousand secrets of the past crawl out of their hiding places – into *his* sunshine."[1] In what follows, wielding the bright light of Lewisian metaphysics, I try to draw into the open some less well-known moments from the history of philosophy. My concern will not be to interpret David Lewis himself, or to reflect on how he may have been influenced by the history of philosophy. Instead, reversing the usual direction of argument, I want to appropriate Lewis's ideas, roughly conceived, as a vehicle for coming to a clearer understanding of some episodes in the history of the discipline.[2]

6.1 Properties

> Any class of things, be it ever so gerrymandered and miscellaneous and indescribable in thought and language, and be it ever so superfluous in characterising the world, is nevertheless a property. So there are properties in immense abundance . . . Because properties are so abundant, they are undiscriminating. Any two things share infinitely many properties, and fail to share infinitely many others. That is so whether the two things are perfect duplicates or utterly dissimilar. Thus properties do nothing to capture facts of resemblance.[3]

For most of the recorded history of philosophy, it has been assumed that an adequate account of language and thought would require postulating *abstracta* of one sort or another, including, inter alia, Platonic Forms, Aristotelian forms, Stoic *lekta*, and Avicennian intentions. To be sure, there were exceptions. After all, one could hardly engage the subject in any serious way without at least pausing to wonder whether it is necessary for the philosopher to traffic in such obscurities. Even so, the massive influence of Plato and especially Aristotle on late antiquity and beyond guaranteed that one or another kind of Realism would dominate philosophy for many a century.

One particularly well-developed line of resistance appears in the later Middle Ages, and was given its most influential statement by William Ockham in the early fourteenth century.[4] Ockham's

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

nominalism stands to analytic metaphysics rather like Darwin's biology stands to natural theology. It is not that Ockham – the "venerable inceptor" of nominalism – put an end to metaphysical realism, but rather that he gave such prominent and articulate voice to the alternative that subsequent discussions could no longer responsibly carry on as they had before. After Ockham, scholastic philosophers self-consciously divided into realist and nominalist camps, and some universities even endowed dual chairs so that each school would have its champion.[5]

For Ockham, nominalism was first and foremost a theory of language. His predecessors had generally assumed that an adequate analysis of subject–predicate statements would require some appeal to a common nature or universal. Ockham, in contrast, argued that language could be explained entirely in terms of concrete particulars: token sentences – spoken, written, or mental – that signify individual things in the world. His general strategy for linguistic analysis is to treat an affirmative sentence as true if and only if its subject and predicate refer to (the technical scholastic term is "supposit for") the same thing or things. On this so-called identity theory of predication, we can say that

(1) Socrates is an animal

comes out true iff "Socrates" and "animal" both supposit, in the context of the sentence, for the same individual. Similarly,

(2) All human beings are animals

comes out true iff "All human beings" supposits for certain individual things and "animals" supposits for all of those same things (it may also supposit for other things). (2) counts as true, then, because the subject-term picks out all the human beings, and the predicate term likewise picks out, among other things, all the human beings. Of course, the theory requires a story about how a term, in the context of a given sentence, and modified by syncategorematic terms like "all," comes to have a certain supposition. But the great advantage of the theory, from Ockham's point of view, is that we need not postulate an ontology of properties inhering in subjects. Speaking of singular affirmative propositions like (1) above, he writes:

> [I]t is not required that the subject and predicate be really identical, or that the predicate be in reality in the subject or that it really inhere in the subject, or that the predicate be united to that subject in extra-mental reality . . . Rather, it is sufficient and necessary that the subject and predicate supposit for the same thing.[6]

What about sentences where the property itself seems to figure as a referent? Even here, Ockham is unrelenting. He holds, for instance, that an abstract term like "animality" refers not to any property but simply to particular animals – different ones depending on the context in which the term appears:

> 'Animality' does not stand for an accident of an animal, nor for one of an animal's parts, nor for any whole of which an animal is a part, nor for any extrinsic thing completely distinct from an animal.[7]

This forces Ockham to accept as true sentences like "Socrates is animality." Better this, though, than admit a world of indefinitely many entities corresponding to all the nominalizations that language could construct from predicates:

> This is the source of many errors in philosophy: to hold that for a distinct word there always corresponds a distinct thing signified, so that the distinctions among things signified match the distinctions among names or significant words.[8]

The cure for this tendency is Ockham's Razor.

Beyond the semantic difficulties that arise from foregoing properties, there are questions about how to explain the similarity between things. It is natural to suppose that Socrates and Fido are similar because they share, among other things, the property of animality. Without any such property, Ockham must either deny the similarity, which would be absurd, or find some other story. His story turns on treating similarity as primitive. Some things just are similar to each other, whereas others are not. Where one finds similarity, there is ultimately nothing more to say than that *This is like that*. The resemblance is ineliminable, and not further analyzable. Accordingly, "God cannot make two white things without their being similar, because the similarity is the two white things themselves."[9]

There is much here for the Lewisian to approve of. The easy allure of *abstracta* has been resisted, replaced by subtle linguistic analysis and the strategic recourse to brute simplicity. But Ockham's nominalism is perhaps too extreme to be defensible. He wants to fashion a semantic theory that can be run not only without universals, Platonic or immanent, but also without individual properties or tropes. He furthermore explicitly excludes any prospect of treating properties as classes of individuals, when he remarks above that animality is not "any whole of which an animal is a part." This precludes the Lewisian strategy of treating animality as the collection of all animals. The costs of this approach are steep, inasmuch as Ockham must struggle mightily to account for many ordinary sentences in natural language.

Ockham's nominalism lies at one extreme on the historical continuum of opinion regarding properties. Subsequent opinion, when it did not reject his views out of hand, tended to soften the austere lines of his approach in one way or another, sometimes by invoking a category of entities known as *modes*, which were understood as something less than real properties but somehow something beyond just substances.[10] Indeed even Ockham himself departed from a perfectly nominalistic rejection of all properties, inasmuch as he recognized certain kinds of individual properties (or tropes) in the category of *Quality*. This, however, raises a new question: how to distinguish between those descriptions of the world that are merely artificial, and those that capture its true nature.

6.2 Carving at the Joints

> The realism that recognizes a nontrivial enterprise of discovering truth about the world needs the traditional realism that recognizes objective sameness and difference, joints in the world, discriminatory classifications not of our own making.[11]

The initial philosophical impulse, scarcely distinguishable from the original scientific impulse, was to discover the fundamental entities that give rise to the world of appearances. The crude efforts of Thales and his contemporaries are familiar enough, as are the more sophisticated efforts of Democritus, who combined anti-realism at the level of appearances with realism at the microscopic level: "By convention sweet and by convention bitter, by convention hot, by convention cold, by convention color; but in reality atoms and void."[12] This sort of reductionism seems to have struck Plato as being too crude even to deserve any mention in his dialogues. It is to Plato that we owe the memorable image of "cutting up each kind according to its species along its natural joints, and trying not to splinter any part, as a bad butcher might do."[13] But Plato is talking about conceptual analysis rather than physical reduction, toward which, as we will see, he was generally hostile.

It was perhaps Aristotle's single most important achievement to fashion a *via media* between the radically opposed programs of Platonic idealism and Presocratic reductionism. The side of Aristotle that leans toward Platonism – the theory of forms – is naturally the most discussed part of the theory,

but from a historical point of view Aristotle's more reductive side is also extremely important. In his work *On Generation and Corruption*, Aristotle embraces – as the best available scientific account – the doctrine of the four elements and their four associated qualities. The elements are Earth, Air, Fire, and Water – capitalized, because, for instance, elemental Earth is not any kind of ordinary earth, but a special chemical element that is found in some proportion or another in every sublunary body. All such earthly stuff arises out of a mixture from each of these four elements. The elements, however, are not actually basic to the scheme. The truly basic explanatory principles are the qualities – Hot, Cold, Wet, and Dry – where again the capitalization stresses that these are theoretical postulates rather than the ordinary sensible qualities. Each element carries with it two such basic qualities,[14] and it is the mixture of these qualities, in varying proportions, that gives rise to the qualitative variation among observable phenomena.

Medieval Aristotelians referred to these four qualities as the *primary qualities*, and put them at the center of their natural philosophy. According to Albert the Great, "the primary qualities of tangible things are the cause of all the other sensible qualities," by which he means the so-called *secondary qualities*, such as color and flavor. Thomas Aquinas says that these four primary qualities are "the cause of generation and corruption and alteration in all other bodies," which is to say that they explain all the most fundamental events in nature.[15] By the seventeenth century, this scholastic terminology was deeply entrenched in the philosophical curriculum, and could be used to plot its downfall. Robert Boyle, turning such Aristotelian vocabulary against itself, contrasts "the primary and mechanical affections, … motion, figure, and disposition of parts" with "those more secondary affections of bodies … which are wont to be called sensible qualities." John Locke, a few years later, would make famous our now-canonical distinction between the primary qualities (by which of course he means size, shape, and so forth) and the secondary qualities (color, heat, etc.).[16]

For a few shining decades, it really seemed as if everything in nature could be explained in terms of geometric–kinetic properties. To be sure, the most careful philosophers of the period did not want to introduce this sort of "mechanical" philosophy as a new philosophical dogma. René Descartes, the most careful of them all, asked the readers of one of his early, unpublished treatises to "allow your thoughts to wander beyond this world to view another world – a wholly new one which I shall bring into being before your mind in imaginary spaces."[17] What he proceeds to imagine is a world without forms and qualities, a world with only solid bodies – *res extensa* – infinitely divisible and put in motion according to the laws of nature familiar from this world. Is that imaginary world in fact *our* world? Descartes, in this early work, was not willing to say so, but he did say that, for all we can tell, it might be our world, inasmuch as a world like that, consisting only of particles in motion, *could* give rise to all the phenomena around us.

Right here, at the start of the "modern" era, we have the Lewisian strategy for thinking about the foundations of reality: admit that the issues are contingent, articulate what looks to be the most elegant theory compatible with the empirical evidence; defend the tenability of that theory against conceptual, philosophical objections. What happened to this Lewisian program in the seventeenth century is that, by the close of the century, it was apparent to well-informed observers that Newtonian forces had spoiled the dreams of a purely mechanical theory. As the young Newtonian John Keill remarked in 1702, "although the mechanical philosophy is today celebrated in name, and in our era its practitioners have attained fame, nevertheless in most of the writings of the physicists one can find hardly anything mechanical beyond the name itself."[18] For the 300 years since then, the metaphysical foundations of science have become steadily more obscure.

Let us then set aside those foundations, and focus on macro-level, ordinary perceptible objects. Here again, from the beginnings of philosophy, one finds doubt about how best to proceed. The same impulses that led Democritus to anti-realism regarding sensible qualities led him to nihilism regarding composition. Plato's disinterest in physical reduction is of a piece with his broader disinterest in

the world of sights and sounds.[19] Again one finds in Aristotle a compromise attempt to save commonsense ontology by deploying forms immanently and inseparably within material objects, as a principle of unity. For as long as Aristotelianism held sway – which is to say for most of the history of philosophy, until around 350 years ago – it was generally supposed that philosophy had a well-understood principle of composition, at least for the paradigm case of living things. (The situation for artifacts was always less clear.) On this approach it is a determinate fact whether, at a given instant, a particular bit of stuff is informed by an animal's substantial form or soul. Accordingly, there is nothing vague about when a substance comes into and goes out of existence, or where its spatial boundaries lie.

Once the Aristotelian consensus collapsed, in the middle of the seventeenth century, the philosophical understanding of substance collapsed with it. Spinoza opts for monism; Hobbes allows unrestricted composition; Leibniz thinks substantial forms must be retained; Descartes seems to have no theory at all.[20] Among the empiricists, the characteristic strategy was to eschew metaphysical speculation in favor of an analysis of our pragmatic interests, as reflected in language and ideas, which are presumed to be divorced from the true metaphysical reality. Thus Locke insists that although he believes in substance as the unifying entity beneath sensible qualities, he thinks this is something we "have no distinct idea of at all."[21] We have, he thinks, no idea of what it is in general to be a substance, nor do we have any idea of what particular substances are, such as a piece of gold or a horse. Still, we cannot escape talking about such things, and so in place of the idea of the real essence that would define a particular substance, we frame the idea of a nominal essence. Thus the way we talk has only a partial connection with the way things are: "the species of things to us are nothing but the ranking them under distinct names according to the complex ideas in us, and not according to precise, distinct, real essences in them."[22]

The era from Descartes to Hume liberated philosophy from the arcana of Aristotelian metaphysics. In so doing, these figures were forced to take seriously the prospect that metaphysics might part ways with common sense. Ultimately, the baroque complexities of scholastic thought served at the behest of a descriptive metaphysics that aimed to do as much justice as possible to our pre-theoretical worldview. But once philosophers tried doing without that marvelous all-purpose device that is the Aristotelian form, they found common sense impossible to save. Of necessity, metaphysics had to become revisionary, or had to be abandoned altogether. Metaphysicians today, Lewisian or not, face much the same choices, pulled in different directions by the comfort of common sense, the allure of speculative metaphysics, and the worry that such speculation is idle.

6.3 Persistence

> A persisting thing is like a parade: first one part of it shows up, and then another. (Except that most persisting things are much more continuous than most parades.) The only trouble with this hypothesis is that very many philosophers reject it as counterintuitive, or revisionist, or downright crazy (except in the case of events or processes). It is a mystery why.[23]

Having slain the Minotaur, escaped the labyrinth, and returned triumphant to Athens, Theseus felt obliged to honor Apollo by sending a yearly tribute to the sanctuary on Delos. This required a ship, and it seemed fitting to use the very same ship that Theseus had used in escaping from Crete. According to Plutarch, that same ship was preserved for centuries in the Athenian harbor, and sent out every year on its religious mission, even down through the time of Aristotle. Of course, a ship that old, moored continuously at sea, required constant maintenance, and it can safely be assumed that, by Aristotle's time, none of the original wood was extant. You know the question.

Puzzlement over persistence goes back to the beginnings of philosophy, and gave rise to protracted debates between those who were skeptical about whether anything persists through change, and those who championed various metaphysical solutions to the problem of change. The sorts of solutions available of course depended on the metaphysical resources that one or another school of thought allowed itself. The Stoics, whose materialism allowed room for an ontology of substances and qualities, individuated bodies both at a time and over time by relying on qualitative sameness and difference. This led them to embrace the identity of indiscernibles.[24] Most, however, assumed that qualitative sameness was quite unsuited to account for identity over time, given the obvious facts about change. Indeed, some did not think that the diachronic identity of substances even allowed for the identity of qualities over time. According to the Asharite school of Islamic theology, only atoms endure through time. The properties that inhere in those atoms, and which give the world its qualitative character, exist only for an instant, and so must be created anew by God at every successive moment.[25] Philosophers in the Latin-Christian tradition were equally quick to develop such surprising views. Peter Abelard and the other *Nominales* of the twelfth century, for instance, commonly endorsed the thesis that "Nothing grows," on the grounds that growth entails a kind of change that is incompatible with sameness.[26]

It might be supposed that Aristotelians are immune to puzzles of persistence, in virtue of having available not only accidental forms, which may come and go, but also the substantial form – for living things, the soul – that individuates substances over time. This may be true for Aristotle himself and for early scholastic Aristotelians such as Thomas Aquinas. Indeed, some scholastics introduced multiple substantial forms within living substances, one in virtue of which it is a body, and one or more others in virtue of which it is an animal or a human being. This allows a single substance to have nested sets of essential properties: to be rational in virtue of its rational soul, for instance, and to be extended in virtue of its bodily form. It is not that there are two things here, a body and a human being, but that there is one complex thing, which in different respects has different essential features, features that might come apart insofar as its different substantial forms might come apart.[27] Again, it can start to feel as if we can use forms to do *anything* in metaphysics.

Beginning in the fourteenth century, however, a certain sort of skepticism arose about whether forms could help at all in solving puzzles of persistence. These worries arose within the later medieval nominalist tradition, and grew out of two principles that Ockham himself had clearly articulated:

Part-Whole Identity. A whole is nothing other than its parts.
The No-Transfer Principle. Forms cannot transfer from subject to subject.

Ockham takes the first principle to entail mereological essentialism, that no whole can survive the loss or gain of any of its parts:

> It is impossible for any one whole thing in its own right, distinct from other things, to exist in reality unless each part of it exists in reality . . . Hence if just one part does not exist in reality, then neither does the whole exist.[28]

This all by itself blocks the simple Aristotelian appeal to substantial form, because even if the substantial form endures through such change, the whole substance cannot endure. But the No-Transfer Principle makes things even worse, because it entails that not even the substantial form can persist through change to a thing's parts. Change at the material level, on this account, forces change at the formal level. Hylomorphism accordingly turns out to be completely worthless in accounting for diachronic identity.[29]

Part-Whole Identity has a long history in these discussions, going back to the twelfth-century *Nominales* and to the Hellenistic debates between the Stoics and the Academic skeptics.[30] For skeptics

regarding persistence, the focus has tended not to be on the Lewisian problem of qualitative change, but rather on the problem of material change – that is, on the gain or loss of integral parts rather than on the gain or loss of properties. This reflects the widespread notion that what endures through time is not the thick substance that is the thing together with all of its properties, but the thin substance, the thing itself, the bare cat shorn of its accidental features. So even though the principle of non-contradiction was generally felt to be binding on all parties, it takes the right sort of contradictory assertions to trigger a violation. That cat can be friendly today and feisty tomorrow, because those are advening states of its soul, not strictly parts of the cat. But if the cat is feisty because it lost a piece of its tail, then that makes for metaphysical trouble. And of course it was common knowledge, then as much as now, that material substances are constantly gaining and losing parts.

Skepticism regarding persistence is a tenable view only when it comes with some sort of further story about why we talk as if things persist. The most famous such account is Locke's, who takes the usual nominalist line regarding material substances, that they endure only for as long as their parts endure: "if one of these atoms be taken away, or one new one added, it is no longer the same mass or the same body."[31] But just as Locke distinguishes between real and nominal essences, so he distinguishes between the strict metaphysical story about substantial persistence and the ideas that figure in how we talk about persistence:

> It is not therefore unity of substance that comprehends all sorts of identity, or will determine it in every case. But to conceive and judge of it aright, we must consider what idea the word it is applied to stands for: it being one thing to be the same *substance*, another the same *man*, and a third the same *person*, if *Person*, *Man*, and *Substance* are three names standing for three different ideas; for such as is the idea belonging to that name, such must be the identity.[32] (1975, *Essay* II.27.7)

It is usually supposed that in passages such as this Locke really means to be talking about true metaphysical identity, and accordingly it is supposed that we must wait until Hume to get a clear account of the difference between an "uninterrupted and invariable object" and a "succession of related objects."[33] But in fact Hume's skeptical line about diachronic identity differs from Locke's mainly in its emphasis. And Locke himself is simply repackaging material that was at the time quite familiar from nominalist discussions. All the way back in the fourteenth century, John Buridan had attempted to make sense, in much the same way, of our casual attitudes toward identity. First, there is strict identity:

> There are three ways in which we are accustomed to say that one thing is numerically the same as another. The first way is by being totally (*totaliter*) the same – namely, because this is that and there is nothing belonging to the whole of this that does not belong to the whole of the other and vice versa. This is numerical sameness in the most proper sense. According to this way it should be said that I am not the same as I was yesterday, for yesterday there was something that belonged to my whole that has now been dissolved, and something else that yesterday did not belong to my whole which later, by nutrition, was made to belong to my whole . . .

Buridan accepts part–whole identity, and accordingly he thinks that numerical sameness is properly had only if a thing retains all and only the same parts. The most that can be said about a human being, then, is that it is partially the same – which is just to say that part of it, the human soul, endures:

> In a second way, however, one thing is said to be *partially* the same as another – namely, because this is part of that . . . And in this way a human being remains the same through the totality of his life because the soul remains totally the same, and the soul is the principal – indeed the very most principal – part. A horse, however, does not remain the same in this way, and indeed neither does the human body . . .

Buridan also accepts the no-transfer principle. Because the human soul is more than just the form of the body – because it is immaterial – it endures through change to that underlying body. In the case of all other material substances, however, their forms depend on their bodies; the ongoing sameness of a horse's soul, for instance, depends on the horse's bodily sameness. Inasmuch as the horse's body is constantly changing, one needs to understand the horse's diachronic identity in a still looser sense, like the identity of a river over time:

> But in a still third way, less properly, one thing is said to be numerically the same as another according to the continuity of distinct parts, one in succession after another. In this way the Seine is said to be the same river after a thousand years, although properly speaking nothing is now a part of the Seine that was part of it ten years ago. For thus the ocean is said to be perpetual, as is this earthly world, and a horse is the same through its whole life and likewise so is the human body.[34]

Such continuity of course does not make for identity in any strict sense, but it explains the sense in which we can truly speak of changing things being the same through time. Locke, more than 300 years later, embraces all of this, and adds to it the brilliant and wholly original suggestion that we form the idea of *personal* identity in still another way, in terms of psychological continuity.

Lewisians will doubtless be heartened by the good sense displayed in these various strategies for retreating from strict identity. Still, they may wonder whether there is room for just one more tiny step, that of calling into question whether anything at all *endures* through time, all of its parts intact, and considering the possibility that instead things might *perdure*, by being composed of a series of ever-changing temporal parts. The question was indeed asked, in the context of debates over *entia successiva*. The idea that there is a divide between two sorts of entities, permanent and successive, goes back to Aristotle. As one example of the familiar principle that being is spoken of in many ways, he offers this: "we say it is day or it is the games, because one thing after another is always coming into existence." The canonical examples are motion and time, about which Aristotle remarks: "time has parts, some of which have been, others of which are going to be, but no part of it is."[35] In later discussions, permanent entities are understood either as those that are capable of wholly existing all at once, or as entities that wholly endure through time. Successive entities fail both of these tests: there can, for instance, be neither time nor motion at an instant, and the whole of time and motion does not endure through its whole existence.

Historically, one finds philosophers of every persuasion embracing successive entities as genuine *things* over and above permanent entities. The idea appears in Aquinas, Arnauld, Augustine, and Avicenna – to canvass just the start of the alphabet. With the concept of an *ens successivum* in mind, it is natural to wonder whether there might be more such things than initially appears, and even whether *everything* might turn out to be an *ens successivum*, on the grounds that nothing endures through time. Nicole Oresme, in the mid-fourteenth century, expressly argued for this possibility, and in particular for the possibility that a rational animal might be created in this way by God, temporal part after temporal part: "such an aggregate from all these would be a human being, a successive substance, of which nothing that existed in a given part of time existed in a subsequent part." Albert of Saxony, building on Oresme's discussion a few years later, argued that this is not just metaphysically but also epistemically possible: that for all we know this is in fact how things are. Still, he adds that he will follow "the general custom" and treat material substances as permanent entities.[36]

So close. But yet so far, and that largely because, for all of the recorded history of philosophy, a peculiar *idée fixe* has held sway, the idea of an enduring material substratum of change.[37] Opinion about the character of this substratum has varied widely. For Democritus and the Epicureans, and again for Gassendi in the seventeenth century, it was atoms that endured beneath all material change. For Aristotelians, it was unformed, "prime" matter. For Descartes it was *res extensa*, infinitely divisible

but essentially extended stuff. Despite dramatic differences in detail, it is hard to find anyone in the history books who did not accept that there is some sort of stuff that endures through all natural change, usually stuff that was thought to be itself, intrinsically changeless. Hence, although substances may come and go, it is never the case that a complete gulf separates what was and what will be. What will be is always composed of some stable ingredients that already were. Hence even when Oresme and Albert of Saxony imagined a successive human being, what they imagined is that God might miraculously do things in this way. They were not imagining that the world might naturally be that way. Nature, or so almost everyone has supposed, for all of the history of philosophy, works by recycling the same enduring ingredients, over and over.

6.4 Causality

> The world has its laws of nature, its chances and causal relationships; and yet – perhaps! – all there is to the world is its point-by-point distribution of local qualitative character.[38]

It would be natural to suppose that, up until the time of Hume, philosophers generally and uncritically accepted the notion of causality in something like our modern sense. In fact, nothing could be farther from the case. The range of pre-modern views is bewildering in the extreme, and Hume's contribution here was not to awaken philosophers from their dogmatic slumbers, but to show off one way in which causality might be analyzed without mystery and obscurity. If Hume's story was hard to believe, at least it posed a salutary challenge: tell me what you think causality is, in a way that is sufficiently credible that I might reasonably rely on that story in framing beliefs about the future.

For the Presocratic atomists, causality was a relatively straightforward affair, a matter of atoms in motion producing motion in other atoms through contact. To be sure, even this very simple picture raises all sorts of hard questions, many of which would become prominent once atomism and allied views returned to center stage in the seventeenth century. But the main course of philosophical thought took a detour that lasted two millennia, a detour that is perhaps largely due to Plato's mockery of reductive mechanistic explanation, most famously in the *Phaedo*:

> When I was a young man I was wonderfully keen on that wisdom which they call natural science, for I thought it splendid to know the causes of everything, why it comes to be, why it perishes and why it exists. I was often changing my mind in the investigation, in the first instance, of questions such as these: Are living creatures nurtured when heat and cold produce a kind of putrefaction, as some say? Do we think with our blood, or air, or fire, or none of these, and does the brain provide our senses of hearing and sight and smell? . . . As I investigated how these things perish and what happens to things in the sky and on the earth, finally I became convinced that I have no natural aptitude at all for that kind of investigation, and of this I will give you sufficient proof. This investigation made me quite blind even to those things which I and others thought that I clearly knew before, so that I unlearned what I thought I knew before . . .[39]

Plato's target is the sort of causal explanations found among Presocratic authors – in effect, the perfectly familiar business of explaining natural phenomena in terms of their basic physical constituents. What such putative explanations distracted him from was something he "knew before," which of course turns out to be the Forms of sensible things. Plato goes on to admit that it may seem naive and foolish to think we can explain beautiful things by their sharing in the Beautiful itself (100d). But the culminating moment of the *Phaedo* is its argument that we can turn our grasp of individual Forms into an account of the interrelationship between distinct Forms, and thereby reach new and substantive conclusions, such as that the soul by its very nature must be immortal.

Although there are vast and contentious questions here for scholars to ponder, what is most salient from a Lewisian perspective is that Plato turns his back on the project of giving explanations that run entirely in terms of local matters of fact. Somehow, particular facts are tied together by a network of necessary connections with higher-order Objects, and these connections are fundamental both to how things are and to how we know them.

When we turn to Aristotle, again we find him seeking a *via media* that, in the present context, amounts to wanting it both ways. Famously distinguishing between four different senses of *cause* or *explanation*, Aristotle wants to leave room for the causal role both of local matters of fact and of necessary connections between those facts. The forms become immanent, and are now conceived of as powers. Powers on their own are not necessarily or inevitably actualized; in many cases they amount to a disposition: the disposition to behave in a certain way, or to enter into a certain state, once the appropriate conditions are realized. In the simplest sort of case, "whenever something capable of acting and something capable of being acted upon are together, what is potential becomes actual."[40] It is often said that efficient causation corresponds to what we now mean by causality, but this is misleading. The efficient cause is simply the agent, and an agent acts (paradigmatically) by bringing its form or power or disposition to bear upon an object that is suited to receive its impression. Efficient causality, then, like material and final causality, depends fundamentally on formal causality. Indeed, Aristotle's ethics, his physics, and his psychology all crucially depend on a theory of forms as immanent powers. Even more fundamentally, his ontology of substance depends on the notion of a governing form or essence that, when realized in an appropriate matter, gives rise to the various defining features of that substance. Again one sees the extraordinary explanatory power of forms, a power that will strike the Aristotelian as a strong recommendation, but may look to the uninitiated as more like theft than good honest work.

Such forms, if they are to be of any value at all, must play an explanatory role in connecting local matters of fact: why certain complex structures persist through time as unified individuals; why certain individuals characteristically behave in certain ways; why certain sorts of behavior is invariably and predictably followed by other sorts of behaviors. The tendency among modern Aristotelians, from Leibniz forward, is to think of these forms in highly abstract, functional terms, as a metaphysical postulate that floats free of the concrete physical story told by science. This was not, however, the later medieval tendency. When scholastic authors defended Aristotle against more reductive approaches, they treated forms, in effect, as a physical hypothesis, an essential ingredient in a complete scientific account of observable phenomena. Phenomena ranging from substantial unity to gravity, and from moral conduct to digestion, were all thought to require forms of one sort or another as an ineliminable part of the causal story. As criticism of the Aristotelian approach gained increasing traction, in the sixteenth and seventeenth century, the tendency of its scholastic defenders was not to treat forms as a strictly metaphysical postulate, as Leibniz later would, but to offer them instead as part of a full scientific account of natural phenomena. Here, the Lewisian might say, we have an honest attempt to make good on the Aristotelian hypothesis. Honest but, unfortunately, empirically discredited.[41]

But if we must give up forms, then what? One solution is to give up entirely on powers and dispositions. This is what Descartes comes close to doing, at least if one sets aside the special case of mind: "there are no powers in stones and plants that are so mysterious . . . that they cannot be explained . . . from principles that are known to all and admitted by all, namely the shape, size, position, and motion of particles of matter."[42] Another way forward was pioneered in England by Robert Boyle and then made famous by John Locke, whose theory of secondary qualities as "nothing in the objects themselves but powers"[43] derives largely from Boyle. Unlike Descartes, Boyle does not seek to get rid of forms and powers. His most important philosophical work introduces as its topic "the nature and origin of qualities and forms, the knowledge of which either makes or supposes the most fundamental

and useful part of natural philosophy." Forms, far from being rejected, remain at the center of Boyle's thought, but they do so merely as explanatory principles that can ultimately be analyzed in terms of the mechanical philosophy. The dissolvability of gold in *aqua regis* is "not in the gold any thing distinct from its peculiar texture"; the poisonousness of the peas "is really nothing distinct from the glass itself." More precisely, the poison is nothing beyond the ground glass contained in the peas, *together with* various anatomical facts about the creature for whom it is poisonous. In general, powers are nothing over and above their categorical bases, but those bases generally extend far beyond the agent that has the power. Thus, "we must consider each body not barely as it is in itself an entire and distinct portion of matter, but as it is a part of the universe . . .".[44]

Our long Platonic detour, hijacked right at the start by the Aristotelians, has now come to an end, leaving us back squarely on the reductive course with which we began among the Presocratics. Still, we are only halfway to Hume, because although we have dispensed with forms and powers as irreducible causal agents, we still have unreduced mechanical impulses – bodies moving other bodies. Locke, even while he accepted such causation, despaired of understanding how it happens: interactions between bodies are "as obscure and unconceivable as how our minds move or stop our bodies by thought."[45] Such worries all by themselves might have produced Hume's doubts over necessary connections, but there is another strand of thought that needs accounting for, the theological strand that shifts the mystery of causation from immanent forms upward into the mind of God. This is, most famously, Nicholas Malebranche's position, whose work was indeed recommended by Hume as preparation for reading the *Treatise of Human Nature*. Some 65 years before Hume, Malebranche had already argued that "when we examine our idea of all finite minds, we do not see any necessary connection between their will and the motion of any body whatsoever. On the contrary, we see that there is none and that there can be none." The same conclusion holds for any two created entities. But Malebranche's conclusion is not that causality involves something other than a necessary connection, but that it involves a different sort of necessary connection, one between God and creatures: "the mind perceives a necessary connection only between the will of an infinitely perfect being and its effects."[46]

Much the same idea occurred centuries earlier, in Islamic thought. According to al-Ghazali, writing at the end of the eleventh century,

> The connection between what is habitually believed to be the cause and what is believed to be the effect is not necessary, according to us. Rather, take any two things. Neither is the same as the other, the affirmation of one does not include the affirmation of the other, and the negation of one does not include the negation of the other. The existence of one does not necessitate the existence of the other, nor does the nonexistence of one necessitate the nonexistence of the other.[47]

Again the point is not to deny necessary connections, but to locate them between God and creatures, and so to identify God as the only true cause. When combined with the earlier-mentioned Asharite denial of enduring properties, the result is a view on which the only things in the material realm that endure are atoms, and on which causation is simply God's creating a new state of the world at each instant. Can this be squared with experience? Here too Islamic occasionalism anticipates Hume's later remark that all we actually *observe* is one event's coming after another. Al-Ghazali, for instance, remarks of his opponent's theory of causality that "their only proof is the observation of the occurrence of the burning upon contact with the fire. But observation proves that the occurrence took place *upon* contact with fire, not that the occurrence took place *by virtue of* contact with fire."[48]

For the occasionalist, the denial of necessary connections, combined with the thesis of empirical equivalence, point toward the rejection of any sort of creaturely causation. The nature of Hume's

own conclusion is less clear. At a minimum, he wants to establish that the idea of a causal power uniting distinct objects can be grounded neither in relations of ideas nor matters of fact, which is to say that this idea has no legitimate support. The only idea we can legitimately have of causation is the idea of constant conjunction, whether that be understood as mere regularity or in counterfactual terms.[49] Arguably Hume wants the stronger conclusion that causation *just is* constant conjunction, but scholars disagree over whether Hume actually intends to go that far.[50]

These are radical conclusions, however exactly they are understood, but they are grounded in familiar, almost uncontroversial premises. The denial of necessary connections between distinct individuals can be found not just in Hume and his occasionalist predecessors, but also in a relatively orthodox Aristotelian account of causality such as Ockham's. It is indeed a fairly obvious point that there are no *logically* necessary connections between distinct individuals. It is equally easy to find precedents for the empirical side of Hume's attack on causality: that observation shows nothing about what causes what. Again, one finds this idea in Ockham, among others.[51] Why then do occasionalists and Humeans derive their startling conclusions from such commonplace principles? The short answer is that more commonsensical views turn out, on reflection, to look deeply unsatisfactory. The Aristotelian's appeal to intrinsic, irreducible powers seems to locative a primitive mystery at the very foundations of the natural world. Powers give rise to connections that are necessary not logically but only in some weaker metaphysical sense: it is of the nature of a thing of a certain kind to act in just such a way, in just such circumstances. Why is this? It just is. As this sort of thinking came to look less credible in the seventeenth century, an alternative account of necessary connections emerged, in terms of natural laws. Francis Bacon dismissively remarks that "forms are fictions of the human soul – unless we are allowed to say that forms are the laws of action."[52] Descartes, even while he rejected Aristotelian powers, eagerly embraced the idea of "certain laws that God has so established in nature, and of which he has implanted such notions in our souls, that after adequate reflection on them we cannot doubt that they are exactly observed in everything that exists or occurs in the world."[53] Isaac Newton was wise enough not to encumber his beautiful physics with metaphysical speculation about what a law might be, but that does not make the philosophical problem go away. Absent some further philosophical account, it is easy to think that the appeal to laws of nature is just a way to get divine influence in by the back door – that we have returned to occasionalism, albeit in a localized, regulated way.[54]

Whether or not Hume himself believed that causation just is conjunction, the difficulties with all rivals views has made the idea look increasingly attractive in modern times. Although a satisfactory analysis in counterfactual terms has proved elusive, it is not clear that historical reflection reveals better options.

6.5 Modality

> I believe that things could have been different in countless ways; I believe permissible paraphrases of what I believe; taking the paraphrase at its face value, I therefore believe in the existence of entities that might be called 'ways things could have been'. I prefer to call them 'possible worlds'.[55]

Though historical generalizations are always hazardous, it seems safe to say that no one before David Lewis attempted to account for modality in terms of real, concrete possible worlds. But even if Lewisian modal realism has no exact historical precedent, there is of course considerable precedent for the idea that possible worlds can somehow contribute to an understanding of modality. We have seen the idea in Descartes already, with his talk of another, imaginary, world where mechanism reigns. Still earlier examples of the usage abound, which is not at all surprising, given how natural it is to speak of "worlds" in describing alternative possible states of affairs.[56]

The most famous case is Leibniz, according to whom "there is an infinity of possible universes in God's ideas" – though he immediately adds that "there cannot exist more than one of them."[57] There cannot exist more than one, because by definition a world is "the entire sequence and the entire collection of all existing things."[58] So only one world exists, but still "there is an infinity" of possible worlds, apparently in the sense that "God's ideas" *represent* infinitely many unrealized worlds. Interestingly, Leibniz denies the possibility of trans-world individuals, and so understands modality in terms of counterpart theory, remarking that when one considers one of the infinity of possible Sextuses, one is not considering "entirely the same Sextus" but rather "a near-Sextus."[59] Leibniz's reasons for this view are idiosyncratic, however, arising not because of how he conceives of possible worlds, but because of his commitment to "superessentialism," which entails that Sextus would not be Sextus if other things had happened to him. In general, "because of the interconnection of things, the whole universe with all its parts would be quite different and would have been different from the beginning, if the least thing in it had happened differently than it did."[60]

How do such appeals to ersatz possible worlds help explain modality? Let us go back much earlier, to some of the first-known accounts of modal language. Diodorus Cronus, circa 300 BCE, is said to have offered the following account:

> The possible is that which either is or will be [true]; the impossible that which is false and will not be true; the necessary that which is true and will not be false; the non-necessary that which either is false now or will be false.[61]

Boethius, our source for this report, immediately complains that this is a hopeless account: it has the absurd consequence that someone who dies at sea could not have died on land. Why would Diodorus have offered something so evidently unsatisfactory? Surely it was not that he failed to recognize more expansive modal notions of the kind that Boethius takes for granted. Our information is so limited that we can only speculate; Arthur Prior has suggested we think of Diodorus as an ancient Quine, offering "some 'harmless' senses that might be attached to modal words."[62] The Lewisian might indeed recognize Diodorus as an ancient master, unwilling to give up modal talk entirely, but resolved to find some solid footing on which to place it, and hence settling for an account that accords only partially with our pretheoretical intuitions. Rather than take the brash Lewisian step of making *possibilia* real, Diodorus settles for limiting *possibilia* to the concrete domain of what is or will occur. (Even this, however, will be less than ideal as solid footing, unless Diodorus is prepared to secure the reality of future events by endorsing eternalism. Until very recently, however, philosophers seem to have almost universally taken for granted that only the present is real.[63])

One finds at least traces of these Diodoran modalities throughout ancient and medieval thought – not because of Diodorus' influence, but because ideas of this kind can be found slightly earlier in Greek thought, in Aristotle. He writes in the *Metaphysics*, for instance, that "it cannot be true to say 'this is capable of being but will not be'."[64] On the usual understanding of how possibility relates to necessity, this entails that what exists always exists necessarily, and in fact Aristotle elsewhere says that "a thing is eternal if it is by necessity; and if it is eternal, it is by necessity."[65] These would seem to be clear statements of what Jaakko Hintikka has labeled a "statistical" model of modality, which analyzes modality in terms of frequency of actual occurrence.[66] It is contentious, however, whether Aristotle actually believes that modality can be *analyzed* in these terms, or even whether he accepts these statistical principles at face value. In his famous discussion of tomorrow's sea-battle, he remarks that "it is possible for this cloak to be cut up, and yet it will not be cut up but will wear out first."[67] This obviously violates the Diodoran rule for possibility, along just the lines that Boethius had complained of. So that rule is surely not Aristotle's. What Aristotle instead seems to have endorsed is a statistical rule for modality in case of things that exist *eternally*.[68] This is exactly what he says, after

all, in the above-quoted passage concerning necessity. And we can understand the bolder passage concerning possibility as limited to cases where a thing is allowed an infinite run of existence. Such a rule is particularly important for Aristotle because he in fact believed that our world, along with the species of things within it, *is* eternal, both in the past and, apparently, in the future. This led him to some surprising ideas about human history. Since he was committed to the view that whatever human beings *could* do is something that in fact they *had done* in the past, he remarks for instance that over the many distant centuries "every art and philosophy has probably often been developed as far as possible and then perished."[69] This looks weird, until one begins to reflect on the implications of saying that human history is literally infinite. And once one begins to reflect on the character of eternity, it might even begin to look as if the *actual* history of *this* world provides a sufficient ground for *possibilia*.

Yet, even if we can make good sense of Aristotle's statistical rule, it seems unlikely that anyone understood it as an account of the grounds of modality. Thomas Aquinas would later be quite clear about this. He dismisses the Diodoran modalities as "*a posteriori*" in the classical sense – that is, as capturing not the nature of modality but at best one if its consequences: "Something is necessary not because it always will be, but rather it always will be because it is necessary, and the same is clear in the other cases." Aquinas then offers his preferred view:

> Others have distinguished these [modalities] better, in terms of the nature of things. On this account, that is said to be necessary whose nature is determined solely to existence; the impossible is determined solely to non-existence; and the possible is wholly determined neither way … This is plainly Aristotle's view here.[70]

The suggestion is that what accounts for modality is not the fact of a thing's occurrence at other times, but the intrinsic nature of a thing. This is another way of making the sort of appeal to forms and powers we observed in the previous section: what a thing can do, and what it must do, are determined by the inner potentialities that determine the web of necessary connections between substances. Again, postulating forms is the easy road to philosophical explanation, and from Aristotle through the Middle Ages, this was the road more traveled.[71]

In the present context, however, it cannot be enough simply to appeal to the forms of actual substances, because this will not capture the full range of possibility, which presumably extends over various unrealized possibilities – natures that have never and will never be instantiated. Here the natural move for theists is to appeal to facts about the nature of *possibilia* as they are in the divine mind. This is a view that one finds running not only through medieval authors, but also into the seventeenth century and beyond. Descartes, for instance, holds that "the eternal truths are true or possible only because God knows them as true or possible. They are not known as true by God as if they are true independently of him." Descartes goes on to identify God's knowing and willing, and adds that "in willing something, he thereby knows it, and thereby alone such a thing is true."[72] This provokes, rather than settles, the familiar Euthyphro-like questions that plague theists: do necessary truths obtain because God wills them, or does God will them because they obtain? Aquinas is more clear about which side he takes: "God *does* something because he wills it; but he *can* do something not because he wills it, but because his nature is so."[73]

With this we can return to Leibniz. Although he is fond of describing modality in terms of possible worlds, in fact it is the divine mind that grounds modality. As we saw, the infinity of possible worlds exists only "in God's ideas." Leibniz is in fact quite explicit about what in fact grounds modal truths:

> Without God there would be nothing real in possibles – not only nothing existent, but also nothing possible. For if there is reality in essences or possibles, or indeed in eternal truths, this reality must be grounded in something existent and actual, and consequently it must be grounded in the existence of the necessary being … [74]

Whereas the Lewisian would force us to choose between admitting real and concerete possible worlds or else denying the Moorean fact of modal truths, the Leibnizian dilemma is either to deny modal truths or admit the existence of God.[75] For Leibniz, talk of possible worlds is simply a *façon de parler*, one that, at least for a modern reader, threatens to obscure the ultimate theological foundations of modality.

What goes for form goes *a fortiori* for God – it makes metaphysics easy, perhaps all too easy. But the theologically-minded have not always supposed that the divine ideas are capable of grounding all modality. A particularly interesting case is John Duns Scotus. He accepts the usual view that one kind of modality is grounded in the powers of things, and another kind grounded in the divine ideas of non-actual things. But Scotus insists on a third kind of modality, for which he coined the term "logical possibility," and which gets spelled out in semantic terms, as obtaining when the terms of a proposition are not incompatible. This idea itself is not new with Scotus. It is found quite expressly in Aquinas as well, who in turn traces the view back to Aristotle.[76] What is interesting about Scotus' discussion (although Aquinas suggests this view as well) is that logical possibility is grounded neither in the powers of actual things, nor in the divine ideas:

> Suppose, before the creation of the world, that there was not only no world but also, *per impossibile*, no God. Suppose that God then began on his own to exist and was capable of creating the world. Then, if there had been an intellect before the world, and that intellect had formed the proposition *The world will exist*, this proposition would have been possible, because the terms are not incompatible. This proposition would have been possible not on the basis of something *in re possibili*, or a corresponding active power, or because of God's potentiality, formally speaking, but because of the potentiality which was the non-incompatibility of the terms.[77]

There is a sense, then, in which we can speak of possibility independently of the natures of what exists, and even independently of God's ideas of their natures.

This new level of modal theory immediately raises the question of what grounds this "non-incompatibility" of terms. The phrase itself, after all, seems to smuggle in modal content. Scotus offers an answer of sorts, in considering why *a human exists* is logically possible but *a chimera exists* is logically impossible:

> The reason *existence* is not incompatible with *human* but it is incompatible with *chimera* is because this is this and that is that – and this holds no matter whose intellect conceives of them.

Evidently, we have hit rock bottom in Scotus' account, with logical possibility turning out to be simply a brute metaphysical fact.[78] One can see the appeal of invoking this further level of modality, but it has the effect of undermining any sense that modal facts have been explained. Lewisians might feel some satisfaction at this result, but might also wonder whether a similar complaint can ultimately be made against even the most boldly realist conception of possible worlds.

6.6 Conclusion

When historians of philosophy tell us that we can learn from the past, they usually have in mind the prospect of discovering neglected ideas that will lead us forward down new and exciting philosophical paths. This can happen. Sometimes, however, the most valuable thing we gain from old philosophy is a proper appreciation of just how few paths forward there are, and how deeply unsatisfactory they look to be. This may cause us to despair. Or it may give us renewed confidence that the paths we are

currently on – difficult, dim, and weird though they may be – are in fact the correct way forward. Or that at least that some of them are. Maybe.

Acknowledgments

Thanks for their considerable help to Dominic Bailey and Bradley Monton.

Notes

1 Nietzsche 2001, §34.
2 On the subject of Lewis's own engagement with the history of philosophy, one should remember his remark *à propos* Leibniz's theory of possible worlds that "It would be nice to have the right sort of talent and training to join in the work of exegesis, but it is very clear to me that I do not" (1986, viii). This remark should in turn be read in light of the fact that Lewis's mother, Ewart Lewis, was a medieval historian.
3 Lewis 1983, reprinted in Lewis 1999, 12–13.
4 Honorable mention goes to Peter Abelard, in the twelfth century, on whose metaphysics see King 2004. One might also think of the Hellenistic era, especially the Epicureans and the Stoics, as early proponents of nominalism.
5 For an overview of later medieval nominalism, see Biard 2010.
6 *Summa logicae* I.6 (William Ockham 1980, 86). For the universal case, see *Summa logicae* II.4. For an overview of the theory, see Panaccio 1999.
7 *Summa logicae* I.6 (William Ockham 1974, 59).
8 *Summulae physicorum* III.7 (William Ockham 1967–89, *Opera phil*. VI.270). For an overview of Ockham's ontological program, see Marilyn Adams (1987, I, 143–67).
9 *Reportatio* II.1 (William Ockham 1967–89, *Opera theol*. V.9). See Marilyn Adams (1987, I, 109–21).
10 See Pasnau 2011, chapter 13.
11 Lewis 1984, in Lewis 1999, 67.
12 The best source for Democritus' texts is Taylor (1999), in which this fragment appears as n. 179.
13 Plato 1997, *Phaedrus* 265e.
14 Air is Hot and Wet. With that hint, the reader will be able to figure out the rest of the combinations. See Aristotle 1984, *Gen. et Cor*. II.1–5, for the overall account.
15 Albert the Great, 1651, *In Gen. et cor*. II.1.1; Thomas Aquinas, 1964, *In De gen. et cor*. prooem. n. 2. For further discussion of medieval views, see Pasnau 2011, chapter 21.
16 Boyle, *Certain Physiological Essays* (1999–2000, *Works* II.98); Locke, *Essay Concerning Human Understanding* (1975, II.8).
17 *The World* ch. 6 (Descartes 1984–91, I.90).
18 Keill 1702, *Introductio ad veram physicam* preface, f. b1r.
19 See Plato's discussion at *Republic* V 476b (1997) of "the lovers of sights and sounds," as opposed to the philosophers, who love Beauty itself. For Democritean nihilism see Aristotle's report: "From these elements Democritus generates the visible and perceptible bodies . . . Entanglement makes them touch and be near one another, but does not really generate any single nature from them; for it would be quite absurd for two or more things ever to become one" (Taylor 1999, n. 44a).
20 For the unfamiliar case of Hobbes, see *De corpore* 11.7 (1905): "a body is always the same, whether the parts of it be put together or dispersed; or whether it be congealed or dissolved." On Descartes' lack of theory, see Pasnau 2011, §25.6.
21 Locke 1975, *Essay* II.23.2.
22 Locke 1975, *Essay* III.6.8.
23 Lewis 2002, 1.

24 Debates over persistence are said to go back to Epicharmus in the fifth century BCE, and make a brief appearance in Plato's *Symposium* (1997, 207d–e). For the extensive Hellenistic debate, between the Academic skeptics and the Stoics, see Long and Sedley 1987, section 28. For the details of the Stoic approach, see Eric Lewis 1995.
25 See Frank 1999 and Griffel 2009, 124–7.
26 See e.g. Arlig 2007 and King 2004.
27 See Pasnau 2011, §25.3. Aquinas, it should be noted, opposed this sort of pluralism regarding substantial form, on the grounds that such composition precludes substantial unity.
28 William Ockham, *In Phys*. IV.18.3 (1967–89, *Opera phil*. V.199).
29 For the details regarding Ockham's view, see Normore 2006 and Pasnau(2011, §29.2).
30 See Barnes 1988.
31 Locke 1975, *Essay* II.27.3.
32 Locke 1975, *Essay* II.27.7.
33 Hume 1978, *Treatise* I.4.6.
34 In *Phys*. I.10 (John Buridan 1964). Amazingly, this powerful and influential work has never been translated into English.
35 John Buridan 1964, *Phys*. III.6, 206a20–30; *Phys*. IV.10, 218a5–8.
36 For references and further discussion of this material, see Pasnau 2011, chapter 18.
37 See Pasnau 2011, §2.1.
38 Lewis 1986, 14.
39 Plato 1997, *Phaedo*, 96a–c.
40 Aristotle 1984, *Physics* VIII.4, 255a34–b1.
41 On the later medieval tendency toward a concrete, physical conception of form see Pasnau 2011, passim.
42 Descartes 1984–91, *Principles of Philosophy* IV.187.
43 Locke 1975, *Essay* II.8.10.
44 *Origin of Forms and Qualities* (Boyle 1979, 2, 24, 25, 26). For further discussion see Pasnau 2011, chapter 23 and Kaufman 2006.
45 Locke 1975, *Essay* II.23.28.
46 *Search after Truth* VI.2.3 (Malebranche 1997, 448, 450). On the relationship between Hume and Malebranche, see e.g. McCracken 1983, Nadler 1996.
47 *Incoherence of the Philosophers* 17, in Khalidi 2005, 159. The view described here had been defended a century earlier by the Asharite theologian al-Baqillani, among others, and it is not clear that al-Ghazali himself is committed to such an austere form of occasionalism. See Frank 1992; McGinnis 2006.
48 *Incoherence of the Philosophers* 17, in Khalidi 2005, 160. Malebranche makes similar remarks (see Nadler 1996, 462).
49 For both of these, in a single sentence, see *Enquiry Concerning Human Understanding* 7.2 (Hume 1902, 76).
50 Proponents of the so-called New Hume make an impressive case that Hume means to be making only an epistemological point about our access to necessary connections between distinct events, but there are also powerful arguments for ascribing to Hume the bolder view. See Read and Richman 2007.
51 See Marilyn Adams 1987, II, 741–58.
52 Bacon 2004, *New Organon* I.51.
53 *Discourse on Method* pt. V (Descartes 1984–91, I.131).
54 For some remarks in this spirit, see Loewer (2012). For historical information, see Milton (1998), who remarks that "by the close of the seventeenth century, the idea that the main objective of natural philosophy lay in the discovery of the laws of nature had triumphed" (I.692).
55 Lewis 1973, 84.
56 For the premodern history of the concept, see Schmutz 2006.
57 *Monadology* par. 53 (Leibniz 1989, 220). See also *Theodicy* (Leibniz 1952) pars. 414–16 – where he speaks of "worlds" (*mondes*) rather than "universes" (*univers*) – and see the useful summary in Parkinson 1995, 212–16.
58 *Theodicy* par. 8 (Leibniz 1952), where uniqueness is expressly derived from the definition.
59 *Theodicy* par. 414 (Leibniz 1952).

60 To Arnauld, May 1986 (Leibniz 1989, 73). The exact details of Leibniz's views here have been the subject of considerably scholarly debate; see e.g. Cover and Hawthorne 1990.
61 Boethius, Second Commentary on the *De interpretatione* (1877–80, II.234).
62 Prior 1967, 16. Also on Diodorus see Denyer 1981.
63 The only defender of eternalism I can find, before our modern era, is John Wyclif, from the late fourteenth century – see Pasnau 2011, 388–90.
64 Aristotle 1984, IX.4, 1047b4.
65 *On Generation and Corruption* II.11, 338a1–3 (Aristotle 1984). See also *Nicomachean Ethics* VI.3, 1139b22–24.
66 See esp. Hintikka 1973, chapter 5.
67 *De interpretatione* 9, 19a13–14 (Aristotle 1984). I am grateful to Dominic Bailey for initially calling my attention to this passage. See also Hintikka 1973, 100.
68 This is the suggestion in Hintikka 1973, 96. For the case of the cloak, see p. 100.
69 *Metaphysics* XII.8, 1074b10–11 (Aristotle 1984); see also *Politics* VII.10, 1329b25–35.
70 Thomas Aquinas 1962, *Commentary on the* De interpretatione I.14 n. 8.
71 There is a large and impressive literature on medieval modal theory. A good place to begin is Knuuttila 1993 and 2012.
72 To Mersenne, May 6, 1630 (Descartes 1984–91, III:24).
73 Thomas Aquinas 1947, *Summa theologiae* 1a 25.5 ad 1.
74 *Monadology* (Leibniz 1989, 43–4).
75 See Robert Adams 1994, chapter 7, with particular attention to why the grounds of modality must be a necessary being with personal attributes that make it appropriately described as *God*.
76 Thomas Aquinas 1947, *Summa theologiae* 1a 25.3c; Aristotle 1984, *Metaphysics* V.12, 1019b22–33.
77 John Duns Scotus 1950–, *Ordinatio* I.7.1 n. 27 (*Opera* IV.118); see also *Ordinatio* I.36 n. 61. Although Aquinas is not nearly so provocative, he might plausibly be read as making the same point at *Summa theologiae* 1a 25.3c (1947) when he introduces his own version of logical possibility with the remark that, if all modality is grounded in God's power, then the doctrine of divine omnipotence becomes either circular or trivial. On the usual reading of Aquinas, however, he makes modality dependent on God; see e.g. Wippel 1981. Avicenna also seems to divorce possibility from God; see Zedler 1976.
78 John Duns Scotus 1950–, *Ordinatio* I.36 n. 60. On logical possibility as a brute fact in Scotus, see King 2001, 193. Scotus's interesting views about modality range over a host of other interesting issues, particularly regarding his rejection of statistical modality, his denial of the necessity of the present, and his embrace of synchronic possibilities, especially in accounting for free will. For an overview see Normore 2003.

References

Adams, M. (1987). *William Ockham*. Notre Dame: University of Notre Dame Press.
Adams, R. (1994). *Leibniz: determinist, theist, idealist*. New York: Oxford University Press.
Albert the Great (1651). *Opera omnia* (21 vols.). Ed. P. Jammy. Lyon.
Aristotle (1984). *The complete works of Aristotle: the revised Oxford translation* (2 vols.). Ed. J. Barnes. Princeton, NJ: Princeton University Press.
Arlig, A. (2007). Abelard's assault on everyday objects. *American Catholic Philosophical Quarterly*, 81, 209–27.
Bacon, F. (2004). *The instauratio magna part II: novum organum and associated texts (The Oxford Francis Bacon* vol. 11). Ed. and trans. G. Rees & M. Wakely. Oxford: Clarendon Press.
Barnes, J. (1988). Bits and pieces. In J. Barnes & M. Mignucci (Eds.), *Matter and metaphysics: fourth symposium hellenisticum*, pp. 223–94. Naples: Bibliopolis.
Biard, J. (2010). Nominalism in the later Middle Ages. In R. Pasnau (Ed.), *Cambridge history of medieval philosophy*, pp. 661–73. Cambridge: Cambridge University Press.
Boethius (1877–80). *In librum Aristotelis peri hermeneias commentarii editio duplex*. Ed. K. Meiser. Leipzig: Teubner.

Boyle, R. (1979). The origin of forms and qualities according to the corpuscular philosophy. In M.A. Stewart (Ed.), *Selected philosophical papers of Robert Boyle*, pp. 1–96. Manchester: Manchester University Press.
Boyle, R. (1999–2000). *The works of Robert Boyle* (14 vols.). Ed. M. Hunter & E.B. Davis. London: Pickering & Chatto.
Cover, J.A. & Hawthorne, J. (1990). Leibniz on superessentialism and world-bound individuals. *Studia Leibnitiana*, 22, 175–83.
Denyer, N.C. (1981). Time and modality in Diodorus Cronus. *Theoria*, 47, 31–53.
Descartes, R. (1984–91). *The philosophical writings of Descartes* (3 vols.). Trans. J. Cottingham, R. Stoothoff, D. Murdoch & A. Kenny. Cambridge: Cambridge University Press.
Frank, R.M. (1992). *Creation and the cosmic system: al-Ghazâlî and Avicenna*. Heidelberg: Carl Winter Universitätsverlag.
Frank, R.M. (1999). The Ash'arite ontology I: primary entities. *Arabic Sciences and Philosophy*, 9, 165–231.
Griffel, F. (2009). *Al-Ghazālī's philosophical theology*. Oxford: Oxford University Press.
Hintikka, J. (1973). *Time and necessity: studies in Aristotle's theory of modality*. Oxford: Clarendon Press.
Hobbes, T. (1905). *The metaphysical system of Hobbes*. Ed. M.W. Calkins. Chicago: Open Court.
Hume, D. (1902). *Enquiries concerning the human understanding and concerning the principles of morals*. Ed. L.A. Selby-Bigge, 2nd ed. Oxford: Clarendon Press.
Hume, D. (1978). *A treatise of human nature*. Ed. P.H. Nidditch, 2nd ed. Oxford: Clarendon Press.
John Buridan. (1964). *Quaestiones super octo physicorum*. Frankfurt: Minerva. (Orig. work 1509.)
John Duns Scotus (1950–). *Opera omnia*. Ed. C. Balić *et al*. Vatican: Scotistic Commission.
Kaufman, D. (2006). Locks, schlocks, and poisoned peas: Boyle on actual and dispositive qualities. *Oxford Studies in Early Modern Philosophy*, 3, 153–98.
Keill, J. (1702). *Introductio ad veram physicam seu lectiones physicæ habitæ in schola naturalis philosophiae Academiae Oxoniensis, quibus accedunt Christiani*. Oxford.
Khalidi, M.A. (2005). *Medieval Islamic philosophical writings*. Cambridge: Cambridge University Press.
King, P. (2001). Duns Scotus on possibilities, powers, and the possible. In T. Buchheim, C.H. Kneepkens & K. Lorenz (Eds.), *Potentialität und Possibilität: Modalaussagen in der Geschichte der Metaphysik*, pp. 175–99. Stuttgart: Frommann-Holzboog.
King, P. (2004). Metaphysics. In J. Brower & K. Guilfoy (Eds.), *The Cambridge companion to Abelard*, pp. 65–125. Cambridge: Cambridge University Press.
Knuuttila, S. (1993). *Modalities in medieval philosophy*. London: Routledge.
Knuuttila, S. (2012). Modality. In J. Marenbon (Ed.), *The Oxford handbook of medieval philosophy*, pp. 312–41. Oxford: Oxford University Press.
Leibniz, G.W. (1952). *Theodicy*. Ed. A. Farrer, trans. E.M. Huggard. New Haven, CT: Yale University Press.
Leibniz, G.W. (1989). *Philosophical essay*. Trans. R. Ariew & D. Garber. Indianapolis: Hackett.
Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.
Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, 61, 343–77.
Lewis, D. (1984). Putnam's paradox. *Australasian Journal of Philosophy*, 62, 221–36.
Lewis, D. (1986). *On the plurality of worlds*. Oxford: Blackwell.
Lewis, D. (1999). *Papers in metaphysics and epistemology*. Cambridge: Cambridge University Press.
Lewis, D. (2002). Tensing the copula. *Mind*, 111, 1–13.
Lewis, E. (1995). The Stoics on identity and individuation. *Phronesis*, 40, 89–108.
Loewer, B. (2012). Two accounts of laws and time. *Philosophical Studies*, 160(1), 115–37.
Locke, J. (1975). *An essay concerning human understanding*. Ed. P.H. Nidditch. Oxford: Clarendon Press.
Long, A.A. & Sedley, D.N. (1987). *The Hellenistic philosophers*. Cambridge: Cambridge University Press.
McCracken, C.J. (1983). *Malebranche and British philosophy*. Oxford: Clarendon Press.
McGinnis, J. (2006). Occasionalism, natural causation and science in al-Ghazālī. In J.E. Montgomery (Ed.), *Arabic theology, Arabic philosophy: from the many to the one: essays in celebration of Richard M. Frank*, pp. 441–63. Leuven: Peeters.
Malebranche, N. (1997). *The search after truth*. Trans. T.M. Lennon & P.J. Olscamp. Cambridge: Cambridge University Press.

Milton, J.R. (1998). Laws of nature. In D. Garber & M. Ayers (Eds.), *The Cambridge history of seventeenth-century philosophy*, pp. 680–701. Cambridge: Cambridge University Press.
Nadler, S. (1996). "No necessary connection": the medieval roots of the occasionalist roots of Hume. *Monist*, 79, 448–66.
Nietzsche, F. (2001). *The gay science*. Ed. B. Williams, trans. J. Nauckhoff. Cambridge: Cambridge University Press.
Normore, C.G. (2003). Duns Scotus's modal theory. In T. Williams (Ed.), *The Cambridge companion to Duns Scotus*, pp. 129–60. Cambridge: Cambridge University Press.
Normore, C.G. (2006). Ockham's metaphysics of parts. *Journal of Philosophy*, 103, 737–54.
Panaccio, C. (1999). Semantics and mental language. In P.V. Spade (Ed.), *The Cambridge companion to Ockham*, pp. 53–73. Cambridge: Cambridge University Press.
Parkinson, G.H.R. (1995). Philosophy and logic. In N. Jolley (Ed.), *The Cambridge companion to Leibniz*, pp. 199–223. Cambridge: Cambridge University Press.
Pasnau, R. (2011). *Metaphysical themes 1274–1671*. Oxford: Clarendon Press.
Plato (1997). *Complete works*. Ed. J.M. Cooper. Indianapolis: Hackett.
Prior, A.N. (1967). *Past, present, and future*. Oxford: Clarendon Press.
Read, R. & Richman, K.A. (2007). *The new Hume debate*, rev. ed. London: Routledge.
Schmutz, J. (2006). Qui a inventé les mondes possibles? In J.-C. Bardout & V. Jullien (Eds.), *Les mondes possibles*, pp. 9–45. Caen: Presses universitaires de Caen.
Taylor, C.C.W. (1999). *The atomists: Leucippus and Democritus*. Toronto: University of Toronto Press.
Thomas Aquinas. (1947–8). *Summa theologiae*. Trans. L. Shapcote. New York: Benzinger.
Thomas Aquinas. (1962). *Aristotle on interpretation: commentary by St. Thomas and Cajetan*. Trans. J. Oesterle. Milwaukee: Marquette University Press.
Thomas Aquinas. (1964). *Exposition of Aristotle's treatise on generation and corruption, Book I, cc. 1–5*. Trans. P. Conway & R.F. Larcher. Columbus, OH: College of St. Mary of the Springs.
William Ockham. (1967–89). *Opera philosophica et theological*. St. Bonaventure, NY: Franciscan Institute.
William Ockham. (1980). *Ockham's theory of propositions*. Trans. A. Freddoso & H. Schuurman. Notre Dame, IN: Notre Dame University Press.
William Ockham. (1974). *Ockham's theory of terms*. Trans. M. Loux. Notre Dame, IN: Notre Dame University Press.
Wippel, J.F. (1981). The reality of nonexisting possibles according to Thomas Aquinas, Henry of Ghent, and Godfrey of Fontaines. *Review of Metaphysics*, 33, 729–58.
Zedler, B.H. (1976). Why are the possibles possible? *New Scholasticism*, 50, 504–21.

7

David Lewis's Place in Analytic Philosophy

SCOTT SOAMES

By the early 1970s, and continuing through 2001, David Lewis and Saul Kripke had taken over W.V.O. Quine's leadership in metaphysics, epistemology, philosophy of language, and philosophical logic in the English-speaking world. Quine, in turn, had inherited his position in the early 1950s from Rudolf Carnap, who had been the leading logical positivist – first in Europe and, after 1935, in America. A renegade positivist himself, Quine eschewed apriority, necessity, and analyticity, while (for a time) adopting a holistic version of verificationism. Like Carnap, he placed philosophical logic and the philosophy of science at the center of philosophy. While not entirely avoiding metaphysics and epistemology, he tried to "naturalize" both. By contrast, Lewis and Kripke embraced the modalities Quine rejected.[1] They also had no sympathy for his early verificationism, or his twin flights from intension and intention. As for philosophy of science, it was transforming itself into specialized philosophies of the several sciences, and did not lend itself to unified treatment. Although Lewis had deep interests in scientific issues, and was commendably realist about science in general, science itself was not the center of his own distinctive approach to philosophy.

Despite similarities in their opposition to Quine, the differences between Lewis and Kripke were large – especially in the semantics and metaphysics of modality. They also had different philosophical styles. Whereas Lewis was a wide-ranging thinker who pieced together a systematic philosophical world-view, Kripke gave little thought to system, focusing instead on a few central topics. There is, therefore, no conflict between the two on many of the issues on which Kripke was silent. However, the modal semantic and metaphysical issues on which they differed were so central to the philosophy of their era that their deep differences had widespread repercussions.

I begin with one such difference, broached in Lewis (1973a).

> I believe that there are possible worlds other than the one we happen to inhabit. If an argument is wanted, it is this. It is uncontroversially true that things might be otherwise than they are … But what does this mean? Ordinary language permits the paraphrase: there are many ways things could have been besides the way they actually are. On the face of it, this sentence is an existential quantification. It says that there

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

> exist many entities of a certain description, to wit 'ways things could have been'. I believe things could have been different in countless ways: I believe permissible paraphrases of what I believe: taking the paraphrase at its face value, I therefore believe in the existence of entities that might be called '*ways things could have been*'. I prefer to call them 'possible worlds'. (84, my emphasis)

Although there is much here with which to agree, there is also a point of contention. There are many ways my desk could have been but isn't; for example, it could have been clean and uncluttered. These "ways" – *being clean* and *uncluttered* – are not themselves *desks* existing somewhere else, numerically distinct from mine; they are *properties* my desk could have had but doesn't. To call them "possible desks" would not be to respect common sense and ordinary language but to flout them. Similarly, the "ways" posited in modal semantics are *not* alternate concrete universes, but *properties* (treated as maximal) that could have been instantiated. Metaphysically possible worlds or, better, *world-states*, are properties that the (or a) universe *could have had*. For a proposition p to be true *at* w is for it such that *had w been instantiated, p would have been true*. For a sentence S to be true at w is for S to (actually) express a proposition that is true at w.[2]

The contrast with Lewis's "modal realism" is illustrated by (1), drawn from Lewis (1986a, 9–12).

(1) Hubert Humphrey could have been so-and-so.

According to the Kripke-inspired view just sketched, (1) is true iff there is a world-state the instantiation of which would have made the proposition *that Humphrey is so-and-so* true. By contrast, Lewis takes (1) to be true only if there is a concrete universe w causally disconnected from ours, and an individual HH who is both *so-and-so* and *an inhabitant of w*.[3] Humphrey inhabits our universe and is, we may suppose, *not* so-and-so. Since HH is, while Humphrey is not, so-and-so, Humphrey and HH are *not* the same person.[4] Rather, HH is Humphrey's *counterpart*, sharing the characteristics of him we wish to keep fixed when evaluating (1).

This is already shocking. How could the truth of the claim that Humphrey could have won the 1968 US presidential election, require the existence of an *alternate universe* in which *someone similar to Humphrey* wins an election similar to the US presidential election of 1968? Surely, if the truth of our original claim did require that, we would be in no position to know or justifiably accept it. Lewis, implausibly, thought otherwise; worse, he took his metaphysical story to be *what we mean* when we say that Humphrey could have won. The "incredulous stare" (as he put it) that was regularly evoked by his claim is evidence that it isn't what we mean.

Lewis thought that incredulous staring should cease, because *any analysis of modal truth* must identify Humphrey's being so-and-so at w with something, potentially other than him, being a constituent of w and bearing some property potentially other than being so-and-so.

> I think counterpart theorists and ersatzers [Lewis's derogatory name for those whose modal metaphysics differs from his] are in perfect agreement that there are other worlds (genuine or ersatz) *according to* which Humphrey . . . wins the election. And we are in equal agreement that Humphrey . . . is not *part* of these other worlds. Somehow, perhaps by containing suitable constituents . . . but anyhow not by containing Humphrey himself, the other world represents him as winning . . . counterpart theory does say . . . that someone else . . . enters into the story of how it is that another world represents Humphrey as winning. . . . Insofar as the intuitive complaint is that someone else gets into the act, the point is rightly taken. But I do not see that is any objection, any more than it would be an objection against ersatzism that some abstract whatnot gets into the act. What matters is that someone else, or the abstract whatnot, should not crowd out Humphrey himself. And there all is well. Thanks to the victorious counterpart, Humphrey himself has the requisite modal property.[5]

Suppose it is false at the actual world(-state) that Humphrey is so-and so. Still, it may be true at w. For this to be so, Lewis argued, it can't be required that Humphrey be both a genuine constituent of w and so-and-so (since he isn't so-and-so). So, Lewis assumed, what is required must be either (i) that something other than Humphrey is both so-and-so and a constituent of w, or (ii) that Humphrey has some property related to *being so-and-so*, while also being, somehow, a constituent of w, or (iii) that something other than Humphrey is a constituent of w that has a property related to *being so-and-so*. (i) is counterpart theory;[6] (iii) is exemplified by an "ersatz view" in which "worlds" are sets of sentences, some expression denoting Humphrey is a constituent of one of the sentences of w, and the property it has *is referring to an individual said by the w-sentences to be so-and-so*; (ii) is a variant of (iii) that Lewis illustrates with the idea that Humphrey might be treated as a word that denotes itself, and *occurs in a sentence that says of its referent that it is so-and-so*.[7] In each case, Humphrey's having the modal property *being one who could have won the election* is constituted either by someone else's having the property *being one who wins the election*, or by something having some other property. Hence, it is no objection to counterpart theory that it views modal properties as indirectly constituted; every theory must.

Really? According to the Kripke-inspired alternative sketched above, Humphrey could have won the election iff some maximal property w, of which he himself is a constituent, could have been instantiated, and if it had been, would have assured his victory. Is this an instance of Lewis's variant (ii), in which the related property is *being such that one would have been so-and-so if w had been* instantiated? Call it what you like. The key point is that the Kripke-inspired analysis does *not* attempt to explain how Humphrey's having the property *being one who could have won the election* is constituted. Instead, it takes the modal notions used to state (1)'s commonsensical truth conditions to be primitive.

This is no more objectionable than using quantifiers to state the truth conditions of quantificational sentences. It isn't a requirement on a semantic analysis of a construction that "the analysis" constitute a reductive definition. Why, then, should we require modal semantics to allow us to *reduce* modal to non-modal notions? Although Lewis had what was for him an important reason for doing so, Kripke and many others regard the escape from Lewis's unacceptable alternatives to be itself a compelling reason not to go down that road. The way to make this escape is *not* to retreat to fictionalism about possible worlds, which is only attractive by comparison with Lewis's unrealistic "realism."[8] The way to escape Lewis's metaphysical jungle is to take certain modal notions to be primitive, and to use them to construct a plausible conception of possible world-states.

For the Kripkean, world-states are properties ascribable to the universe. Just as we recognize properties ordinary objects don't have, but could have had, so we should recognize (maximal) properties – metaphysically possible world-states – the universe doesn't have, but could have had. For any such property w, the universe *could* have had w, and if it did, certain propositions *would have* been true. Note the use of modal notions – *what could have been*, and *what would have been true if such and such had been so-and-so*. Possible world-states are not defined in non-modal terms. Nor is there any attempt to reduce these two modal notions to non-modal notions. Nothing rules out such a subsequent reduction, but nothing requires it either.

The chief reason Lewis couldn't, or wouldn't, adopt this non-reductionist approach came from his commitment to an ambitious version of philosophical naturalism, the goal of which was to reduce the mental to the physical, causation to counterfactuals, counterfactuals to possible worlds, and possibility and necessity to quantification over ordinary (though very large) objects. The desired destination was a homogeneous reality consisting of momentary point-sized objects (given in physics) instantiating purely qualitative universals – natural properties and relations (also given in physics) – related by similarity, difference, and spatiotemporal relations across multiple universes, but *not* related by "occult," empirically unexplainable forces or connections. To those attracted by this vision,

one may say, "Wishing doesn't make it so." For the faithful, however, the final reduction (of necessity and possibility to quantification over the large but ordinary) is indispensable, since without it the other reductions don't lead to any exciting or satisfying system.

Here we see what appears to have been the enduring influence of David's teacher, Quine, the great champion of naturalism and extensionalism, and the uncompromising scourge of the modalities. The underlying philosophical purpose of modal realism and counterpart theory was to reduce an intensional object-language to a purely extensional semantic metalanguage, in the service of an antecedently desired conception of reality. Whereas Quine taught that vindicating naturalism and extensionalism required eliminating intensional facts and rejecting intensional constructions, his student, David Lewis, tried to show that intensional facts are just a species of extensional facts, and that intensional constructions in language are no threat to the integrity of an austere, naturalistic vision of reality.

With this in mind, consider the following passage from section 45 of *Word and Object*. "If we are limning the true and ultimate structure of reality, the canonical scheme for us is the austere scheme that knows no quotation but direct quotation and no propositional attitudes but only the physical constitution and behaviour of organisms."[9] Although Quine's immediate topic was propositional attitudes, the intent of his remark went much further: no indirect quotation, no operators on *that*-clauses (including 'it is possible/ necessary that'), austere extensional systems of representation, and, in the reality to be described, only the physical constitution of things and the behavior of organisms. Whereas Quine married metaphysical naturalism to overtly extensional systems of representation, Lewis strove to save the former by divorcing it from the latter, using extensional descriptions of a metaphysically stupefying reality to interpret superficially intensional systems of linguistic representation.

The parallel between Lewis and Quine was not limited to their metaphysical outlook; it also included their mature views of physics. After his marked holistic verificationist period, the author of Quine (1969, 1970) wrote mostly as a scientific realist, as illustrated in a passage from the former.

> theory in physics is an ultimate parameter. There is no legitimate first philosophy, higher or firmer than physics, to which to appeal over physicists' heads . . . [A]dopt for now my fully realistic attitude toward electrons and muons and curved space-time, thus falling in with the current theory of the world despite knowing that it is in principle methodologically under-determined. Consider, from this realistic point of view, the totality of truths of nature, known and unknown, observable and unobservable, past and future. The point about indeterminacy of translation is that it withstands even all this truth, the whole truth about nature. (Quine 1969, 303)[10]

For Lewis, as for the later Quine, physical science was the ultimate parameter, which no philosophical theory could afford to contradict, radically reinterpret, or diminish the epistemic and metaphysical significance of. Unlike Russell (1914, 1918–19), or Carnap (1928), Lewis did not favor any attempt to translate or reinterpret science in phenomenalistic terms. Unlike Quine (1948, 1951) or Carnap (1950), he never dreamed of identifying the content of a scientific theory with the totality of observational statements supporting it. Unlike Van Fraassen (1980), he didn't merely *accept* physical theory as way of producing reliable observational results; he believed it, and took it to be our most accurate conception of reality.

In keeping with his scientific realism, Lewis didn't attempt to explain away theoretical terms, but took them to stand for genuine realities. For him, their referents were determined by Ramsification; electrons and muons, for example, are whatever elements of reality best fill the 'electron' and 'muon' roles in our theories (making them at least approximately true). In this he diverged from Kripke. For the latter, many theoretical terms – including 'heat', 'light', and 'electricity' – refer to natural kinds

with which we are cognitively acquainted (through their instances) in a way that secures reference, without tying it to the approximate truth of any theory in which the terms occur.[11] While versions of Kripke's story have won widespread acceptance (for many such terms), it is not unreasonable to suppose that some highly theoretical terms may best be understood roughly along the lines Lewis suggests.[12] Kripkeans needn't have a problem in accepting a version of this limited friendly amendment. Whether or not followers of Lewis can afford such eclecticism is less clear. His own commitment to descriptive reference fixing was nothing if not ferocious and pervasive. In addition to covering all natural kind terms and ordinary proper names, it was also embedded in his thoroughgoing analytic functionalism.[13]

Returning to Lewis's audacious attempt to reduce the modal to the non-modal – as the capstone of the reductions implementing his grand naturalistic program – I note his parallel treatment of temporal modification. What is it for o to have been *so-and-so* at a past time t (where *being so-and-so* is an intrinsic property o now lacks)? As in the modal case, Lewis thinks it must be for some object different from, but related to, o to be both *so-and-so* and located (i.e. existing) at t. Thus, he takes it to be incoherent to suppose that o might itself bear different (intrinsic) properties at different times, in the same way he takes it to be incoherent to suppose that o itself might (without the help of counterparts) bear different such properties at different possible worlds. The chief difference is that the temporal counterpart of o is a time slice of o, while o is itself supposed to be a parade of time slices. By contrast, the modal counterpart of o is an entity on a par with, but non-identical to, o.

As before, this move from mundane semantics to revisionary metaphysics is, I think, unfortunate. The first step in avoiding it is to trade the time-relativized object for the time-indexed property *being so-and-so at t*. But, Lewis rightly objects, that property is relational rather than intrinsic, just as *being so-and-so at w* is. The key to overcoming this objection is to define the world-state and time-indexed properties in terms of the intrinsics (rather than vice versa). For o to be *so-and-so at* w is for it be the case that *if w were instantiated, then o would have the property so-and-so*. For o to be *so-and-so at t* (where t is a past time) is for it to have been true when t occurred that *o had the property so-and-so*. Temporal semantics does not, contra Lewis, require time-slices.

Viewing world-states as properties the instantiation of which *would* make various propositions true, Kripkeans don't take counterfactuals to be *defined* by, or reduced to, conceptually prior world-states. This may seem to flout the orthodox view of counterfactuals in Lewis's (1973a). However, the appearance of heterodoxy is exaggerated; much of Lewis's influential treatment of the truth conditions of sentences of the form (2) can be accepted.

(2) If it had been the case that A, then it would have been the case that B ($A \mathrel{\Box\!\!\rightarrow} B$).

His idea was that (2) is true at w iff either there are no A-states (world-states at which A is true), or some A-states at which B is true are more similar to w than any A-states at which B is false. His semantics differs from one developed independently in Stalnaker (1968) in not validating conditional excluded middle – $(A \mathrel{\Box\!\!\rightarrow} B) \vee (A \mathrel{\Box\!\!\rightarrow} \sim B)$.[14]

What is it for one world-state to be similar to another? Lewis doesn't say much about this in *Counterfactuals* (1973a), beyond noting that what counts as similarity can vary from one context of utterance to another. Challenged by counterexamples to the analysis when similarity is understood to be *overall similarity* of the world-states,[15] he responded in Lewis (1979b) with default rules for determining similarity. Since then there has been much back and forth concerning what relation S_C is needed to complete the account of counterfactuals. Although there is reason to believe that Lewis originally did think that some kind of qualitative similarity was central to the semantics of counterfactuals (just as it was to his counterpart relation), the subsequent literature, including his own later contributions, make it clear that the fact that S_C was originally called "similarity" was merely a distraction.

At this point, we do well to reflect again on the grand naturalistic reduction for which Lewis was striving. In addition to *reducing* possible worlds to other universes, it required *reducing* counterfactuals to possible worlds plus similarity. But once simple qualitative similarity went by the boards, there was no real *analysis* on offer, but only the relatively innocuous statement that in order for it to be true (at w) that *if A were true, then B would be true*, B must be true at those worlds (if any) at which A is true which also bear some further unspecified relation to w. Whatever its merits, this is no definition, analysis, or reduction.

The next stage in the saga – finding the relation S_C needed to complete the "analysis" of counterfactuals – produced both genuine progress and a further threat to Lewis's grand plan. Recent work – including Jonathan Bennett (2003), Dorothy Edgington (2004), Jonathan Schaffer (2004), and Boris Kment (2006) – suggests that similar facts in $w_1$and w_2 contribute to the worlds being closely related by S_C, only if the facts have the same, often causal, explanation in both. Although this extends Lewis's story, and provides needed content, it also raises a worry. Which, if either, of causation and counterfactual dependence is to be analyzed in terms of the other? According to Lewis (1973b), causation is defined in terms of *causal dependence*, which is a form of *counterfactual dependence*. Roughly put, an event b *causally depends* on a distinct, non-overlapping event a iff *if a were to occur, then b would occur* (and if a were not to occur, then b would not occur). A* is *a cause* of a distinct, non-overlapping event b* iff b* is causally dependent on A*, or there is an event c* which both causes b* and is causally dependent on a*.

In the decades after this analysis appeared, many objections were raised, leading Lewis to further modifications – the last of which are found in Lewis (2004). But the direction of analysis remained constant; causation is analyzed in terms of counterfactuals. This is threatened by the idea that causal explanation might itself be needed to explain the S_C-relation in terms of which counterfactuals are to be "analyzed." If, as Lewis intended, these analyses are to express conceptual and explanatory priorities, then, either counterfactuals are *not analyzable* in terms of S_C, or S_C is *not analyzable* in terms of the causal features of worlds, or causation is *not analyzable* in terms of counterfactuals – or all three. As I write, the jury seems to be out on this. Perhaps, the best strategy is to acknowledge interdependence among these, even though to admit this is to give up the idea of *reducing* causality to anything involving counterfactuals.

Such a loss would be no minor matter. In the context of Lewis's grand naturalistic reduction, the role of his counterfactual analysis of causation was to solve Hume's problem by reconciling (i) the way in which our thought and talk of causation is embedded in counterfactual reasoning with (ii) an ultimately Hume-like constant-conjunction analysis of causation – in terms of spatiotemporal relations among events involving point-sized counterpart time-slices instantiating natural properties across universes.[16] If, however, there is no such payoff in the offing, then there will be no philosophical treasure to pay the high price of Lewis's modal realism, and his embrace of counterparts and temporal parts as explanatorily basic.

Earlier, in outlining the structure of his grand naturalistic project, I mentioned the reduction of the mental to the physical. The *locus classicus* for this reduction is Lewis (1980), the precursor of which was Lewis (1966). In both, he self-consciously adopted a version of the contingent identity theory of the relation between mind and body. By 1980, this implicitly involved a counterattack against Kripke's celebrated attempt to refute such theories. To understand Lewis, it is helpful to understand the Kripkean position it repudiates.

Kripke's discussion in *Naming and Necessity* focuses on (3) and (4).

(3) Heat = mean molecular kinetic energy

(4) Pain = C-fiber stimulation

We here treat the expressions flanking '=' as designating kinds and the statements as ordinary identities. Initially, both (3) and (4) appear to be contingent, if true. However, Kripke argues, with (3) this impression is illusory. Suppose that 'heat' and 'mean molecular kinetic energy' are rigid designators – that is, that the state that is heat couldn't have existed without being heat, and similarly for the state of having such-and-such mean molecular kinetic energy. Then, it is thought, (3) will be necessary, if true. Since it *is* true, it is also necessary.

What, then, is responsible for the illusion that it isn't? Kripke's answer focuses on how we identify or "fix the referent of" 'heat'. Since our primary means of identifying heat is by the sensations it causes, he imagines us using the description 'the cause of such-and-such sensations'. The illusion that (3) is contingent comes from mistaking this (non-rigid) description for a synonym of 'heat', thereby confusing the necessary truth expressed by (3) with the contingent truth expressed by (3*).

(3*) The cause of such-and-such sensations in us = mean molecular kinetic energy

One who makes this mistake wrongly takes genuinely possible world-states at which kinetic energy exists without the usual accompanying sensations to be world-states with kinetic energy but no heat. Hence, the illusion of contingency.

Kripke finds (4) to be different. As before, he takes the terms flanking '=' to be rigid, in which case (4) is necessary, if true. However, he argues, there is no way of dismissing the initial impression of contingency. With (3), the impression was (allegedly) caused by our reliance on sensations to detect what causes them. With (4), the sensation itself is what we use 'pain' to talk about. We don't think: *What a horrible sensation! Let's use 'pain' to talk about whatever causes it*. Instead, we use the word to designate the sensation itself. Thus, whereas one can dispel the illusion that (3) is contingent by appealing to the contingent connection between heat and the sensation by which we recognize it, one can't dispel the impression of contingency of (4) that way. Since he sees no other way to do so, Kripke concludes that (4) *must be contingent, if it is true at all*. But, it can't be contingent, since its terms are rigid. So, he suggests, it isn't true.

This argument, which fails to distinguish epistemic from metaphysical possibility, is unconvincing. Although Kripke may have identified one source of the impression that (3) is contingent, he neglected another. Imagine a man who doesn't take 'heat' to be synonymous with any term for our sensations, because he recognizes that heat could exist without us. He might still be under the impression that (3) is contingent because it is *conceivable* for heat to be something other than the motion of molecules. After all, he may reason, it was an empirical discovery that how hot something is depends on how fast its molecules are moving. Since empirical evidence was needed *to rule out possibilities in which this is not so*, it must be *possible* for heat *not* to be molecular motion. Hence, he may (wrongly) conclude, (3) must be contingent. The same reasoning applies to (4). The error is in confusing epistemic with metaphysical possibility. Once it is corrected, the flaw in the argument becomes obvious. Having dismissed only one of two sources of the impression that (4) is contingent, if true, Kripke is not entitled to his conclusion that it's not true.[17]

This objection to Kripke, relying on sharply distinguishing epistemic from metaphysical possibility, would not have appealed to Lewis. For him, contingent theoretical identities involving natural kind terms are unproblematic.[18] On his analysis, pain, for an organism x, is the physical state which, in *normal* members of x's kind, plays a certain functional role; it is typically caused by injuries, it interferes with one's activities, and it is something the avoidance of which motivates action. The theory, articulated in Lewis (1980), is designed to accommodate normal pain plus "mad" and "Martian" pain.

Mad Pain
There might be a strange man . . . whose pain differs greatly from ours. . . . Our pain is typically caused by cuts, burns, pressure, and the like; his is caused by moderate exercise on an empty stomach. Our pain is generally distracting; his turns his mind to mathematics, facilitating concentration on that but distracting him from anything else. Intense pain has no tendency whatever to cause him to groan or writhe. . . . He is not in the least motivated to prevent pain or get rid of it. In short, he feels pain but his pain does not at all occupy the typical causal role of pain.[19]

Martian Pain
Also, there might be a Martian . . . whose pain differs greatly from ours in its physical realization. His hydraulic mind contains nothing like our neurons. Rather, there are varying amounts of fluid in many inflatable cavities, and the inflation of any one of these cavities opens some valves and closes others . . . When you pinch his skin you cause no firing of C-fibers – he has none – but rather you cause the inflation of many smallish cavities in his feet. When these cavities are inflated, he is in pain. And the effects of his pain are fitting: his thought and activity are disrupted, he groans and writhes, he is strongly motivated to stop you from pinching him . . . In short, he feels pain but lacks the bodily states that either are pain or else accompany it in us.[20]

While "Martian pain" is easily imaginable, "mad pain" isn't. Is it really possible? Here, we must distinguish epistemic from metaphysical possibility. Let's say that p is epistemically possible iff we can't know *a priori* that p isn't true. Since the connection between neurological states and behavior is *a posteriori*, it is *epistemically possible* for one to be in the neurological state we are in when feeling pain, while having the madman's behavior and motivation. It is also epistemically possible to be in that state and not *feel pain*. However, when one adds that the madman *feels pain*, despite lacking normal *pain behavior* and *pain-avoiding motivation*, it is natural not only to doubt that *mad pain* is metaphysically possible, but also to question whether it is epistemically so. It is conceivable, and presumably possible, for agents to exhibit *some* behavioral and motivational responses to pain that differ from the norm. But this doesn't mean that the wholesale difference exhibited by Lewis's "madman" is either possible or coherently conceivable. By contrast, the epistemic possibility of Martian pain is beyond doubt. As for its metaphysical possibility, caution is in order. While it is very plausible that actual and possible organisms can feel pain despite having physiologies quite different from ours, the limits of these differences are unknown, and likely to fall well short of what we can pre-theoretically conceive. Presumably, however, Lewis could accept this.

He disagrees with Kripke about 'pain' being a rigid designator. Since most general terms are rigid (whether or not they designate natural kinds), his claim that 'pain' isn't rigid is counterintuitive.[21] Could *pain* have existed without being *pain*? Could anything other than *pain* have been that state? Like (5), but unlike (6), the answers to these questions seem to be 'no'.

(5) Could the color *blue* have existed without being (the color) *blue*? Could anything other than *blue* have been that color?

(6) Could the color of a cloudless sky at noon have existed without being the color of a cloudless sky at noon? Could anything other than the color of a cloudless sky at noon have been the color of a cloudless sky at noon?

One way out is to take 'pain' to be equivalent to 'the state of an individual that plays such-and-such functional role'. To do so is to lose the madman while gaining rigidity – whether or not different physical states *realize* pain states (at different times/worlds). True, one loses something, since pain

may no longer be *identified* with any one physical-state type among all those that may realize it. True also, Lewis would have regarded this as insufficient for his ambitious naturalistic reduction. By now, however, it is not clear that this is a serious worry.[22]

Reviewing my discussion of the core components of Lewis's grand naturalistic reduction – of the mental to the physical, of causation to counterfactuals, of counterfactuals to the similarity relation on worlds, and of worlds to existing physical universes – I have found them all to fall short. They are not genuine reductions of one class of things to another. Nevertheless, all but the last (the modal to the non-modal) contained much of value that moved the discussion of the issues with which they were concerned forward. Even the grand reductive attempt was not without value, being the first real effort by a leading figure of analytic philosophy to construct a comprehensive view of the whole of reality since Russell (1918–19) and Carnap (1928).[23]

Because metaphysics was the center of Lewis's philosophy, this mixed verdict is not a happy one for such a gifted thinker. There were, however, other projects to which he contributed much more successfully. His first book, *Convention* (1969), made an important contribution to the study of equilibriums arising as solutions to certain kinds of coordination problems and multi-agent games. His work on conditional probability and its distinctness from the probability of any conditional sentence, in Lewis (1976), was path breaking.[24] He was also one of the leading contributors to the grand project to which many analytic philosophers have devoted themselves – moving the study of language, information, and cognition from unsystematic philosophical speculations to a (still future) stage in which solid and comprehensive frameworks for genuine scientific investigation are in place. It is to this broad project that I now turn.

Lewis's (1970b) was one of the cutting-edge texts of its time – along with work by Richard Montague, David Kaplan, and Robert Stalnaker. Together, they laid out a powerful framework for the use of intensional logic in semantic theories of natural language. In his paper, Lewis used a categorial grammar to generate deep structures, mappable onto surface structures by Chomsky-like transformations (in a manner then being investigated by his colleague Barbara Partee at UCLA). Deep structures were interpreted by a version of Montague's one-semantic-rule-for-each-syntactic-rule approach. The resulting system assigned intensions (functions from indices to truth values) to sentences.[25] Although his system was both powerful and fruitful, it was also resolutely intensional, with no room for hyperintensional operators. In keeping with the assumptions of the time, propositions were identified with intensions of sentences – which, for our purposes, we may take to be functions from "worlds" to truth values, or, more simply, sets of "worlds." Though many today still adhere to this approach, others have come to see it as incorporating limitations that Lewis never transcended, as demonstrated by his discussion and individuation of propositions in such late works as Lewis (1996).

Historically, the three main attempts to deal with hyperintensionality in intensional semantics have been (i) to substitute finer-grained circumstances for world-states, (ii) to distinguish the proposition semantically expressed by S from what one asserts by uttering S, and (iii) to adopt a 2D semantic theory which associates *pairs* of coarse-grained propositions with sentences. The high-water mark for (i) (never favored by Lewis) was the system of situation semantics developed by Barwise and Perry (1983). It failed because the hyperintensionality problem is reconstructable for all theories of propositions as sets of truth-supporting circumstances (satisfying certain minimal assumptions).[26] The high-water mark for (ii), pioneered by Stalnaker (1978) and endorsed by Lewis, identified propositions with sets of metaphysically possible world-states, while combining the distinction between semantic and assertive content with a 2D strategy of "diagonalization" to breathe informativeness into utterances of sentences that express necessary consequences of propositions already in the context. It fails for various reasons, including its inability to accommodate attitudes to singular propositions that predicate essential properties of things.[27] The high-water mark for (iii) is a semantic view, suggested by David Chalmers (1996) and Frank Jackson (1998), called "strong

two-dimensionalism" in Soames (2005). The basic idea is that sentences express propositions dubbed "primary" and "secondary" intensions, the former being the arguments of hyperintensional operators and the latter being the arguments of modal operators. Since the primary intension of S says, roughly, that S's Kaplanian character expresses a truth, primary and secondary intensions diverge when S contains indexicals or indexically rigidified descriptions (which names and natural kind terms are taken to be).[28] The chief technical problem is that strong 2D can't account for sentences containing both modal and hyperintensional operators, one embedded under the other. Rescue attempts that appeal to weaker versions of 2D semantics suffer from similar shortcomings.[29]

Though Lewis didn't himself pioneer this failed approach, he did embrace it.[30] Moreover, the problems with propositions as sets of "possible worlds" are not limited to coarse-grained nature of the latter. It is crucial to any account of propositions that, as bearers of truth conditions, they represent the universe as being certain ways (without requiring interpretation by us). No set is inherently representational in this way. What does the set containing "worlds" 1, 2, and 3 represent? Is it true or false? These questions make no sense. If we wanted, we could use the set to represent the "actual world" as being in the set – and so to make the claim that no "world" *outside* the set is actual. But we could equally use it to represent the "actual world" as *not* being in the set – and so to make the claim that no "world" *inside* the set is actual. Independent of interpretation by us, the set doesn't represent anything, and so has no truth conditions.[31]

This brings me to the necessary *a posteriori*, where several fault lines dividing Lewis and Kripke meet.[32] No theory identifying necessarily equivalent propositions can distinguish necessary propositions that are knowable *a priori* from those that aren't. If I am right, there is also no acceptable 2D-reanalysis of Kripke's data supporting the necessary *a posteriori*. Lewis's treatment of metaphysically possible worlds as *concrete universes* rather than *maximal properties* exacerbates the problem. The alternative treatment, as maximal properties, simply generalizes to the universe the platitude that some things could have had properties they don't have. The further platitude that sometimes a property something couldn't have had can be known not to be had by it *only by appeal to empirical evidence*, allows a further extension. In addition to including maximal properties the universe could have had but doesn't (metaphysically possible world-states), the analysis can accommodate those the universe couldn't have had, but can't be known *a priori* not to have (metaphysically impossible but epistemically possible world-states). With such states, the function of evidence needed to know an empirical proposition p can be just what Lewis (1996) says it is – namely, to rule out relevant *possibilities* in which p is false – even when there are no *metaphysical possibilities* to rule out because p is necessary.[33] By contrast, when one thinks of possible worlds as concrete universes (that exist but don't "actually exist") the idea of metaphysically impossible, but epistemically possible, worlds will, wrongly, seem incoherent.

Mention of actuality highlights another problem with modal realism. Most philosophers agree that ⌈Actually S⌉ is necessarily true when S is contingently true. Less widely recognized is the fact that assertively uttering either sentence typically commits one to the propositions expressed by both. Lewis's counterpart theory prevents him from recognizing these things. Although Hubert Humphrey's counterpart, HH, exists and is so-and-so, according to Lewis he doesn't *actually exist* and is not *actually so-and-so* (because he doesn't exist and is not so-and-so *in our universe*).[34] The right result – that anything that exists and is so-and-so, *actually exists* and *actually is so-and-so* – is readily explainable when 'Actually' is taken to indexically refer to the *world-state* of the context, and ⌈Actually S⌉ is seen as equivalent to ⌈At this very world-state S⌉.[35]

As I have emphasized, Lewis was a systematic philosopher. Sometimes I wish he weren't quite so. One of his most influential papers, Lewis (1979a), contains insights that must be separated from the modal realism that deeply informs it. The insights concern a problem about "first-person" (*de se*)

attitudes previously raised and addressed in Hector-Neri Castaneda (1966, 1968) and John Perry (1977, 1979). Here is Lewis's statement of the problem.

> Consider . . . two gods. They . . . know every proposition that is true at their world . . . Still . . . [n]either one knows which of the two he is . . . One lives on top of the tallest mountain, and throws down manna; the other lives on top of the coldest mountain and throws down thunderbolts. Neither one knows whether he lives on the tallest mountain or on the coldest mountain, nor whether he throws manna or thunderbolts. Surely their predicament is possible. (The trouble might perhaps be that they have an equally perfect view of every part of their world, and hence cannot identify the perspectives from which they view it.) *But if it is possible to lack knowledge and not lack any propositional knowledge, then the lacked knowledge must not be propositional.* If the gods came to know which was which, they would know more than they do. But they wouldn't know more propositions . . . Rather . . . [o]ne . . . would correctly self-ascribe the property of living on the tallest mountain. (1979a, 139, my emphasis)

The gods know that the manna-thrower lives on the tallest mountain, and the thunderbolt-thrower lives on the coldest mountain. Each also knows, of the manna-thrower M, *that he lives on the tallest mountain* – that is, each knows the singular proposition expressed by 'x lives on the tallest mountain' relative to an assignment of M to 'x'. The same is true of the singular proposition expressed by 'x lives on the coldest mountain' relative to an assignment of the other god, T, to 'x'. Nevertheless, they do seem to lack knowledge. If M or T were to say "I don't know whether I live on the tallest/coldest mountain" we would judge him to have spoken truly. This suggests *there are some things they don't know*. Since there are no propositions they don't know, Lewis concludes that not all knowledge is propositional; sometimes what is known is a property that one could correctly self-ascribe. The main alternative at the time was Perry's, according to which the gods already know all there is to know; they simply fail to recognize certain known propositions when presented to them in a first-person way. I believe Lewis is closer to the truth.

Still, there are challenges to be met. The first is illustrated by Perry's (1977) example of the amnesiac Lingens, who is trapped in the Stanford library reading a fact-filled biography of Lingens that includes a description of his predicament. Although Lingens believes, and even knows, all the propositions learned from reading, we naturally describe him as "not knowing that *he* is Lingens," and "not knowing that *he* is trapped in the Stanford Library." Which propositions are these? Not the propositions *that Lingens is Lingens* and *that Lingens is trapped in the Stanford Library*. Everyone who has heard of Lingens knows the former, and Lingens himself knows the latter because the book told him. Despite this, Lewis maintains, Lingens is in no position to self-ascribe *being Lingens* and *being in the Stanford library*. So far, so good. Now consider an extension in which Lingens looks in the mirror and says "That's me!," self-ascribing *being him* (demonstrating himself). Since this *is* the property *being Lingens*, it seems we should conclude that Lingens now *knows that he is Lingens*. But we wouldn't naturally say that. Is Lewis's account therefore incorrect, or unmotivated, because it allows us to recreate a version of the very problem it was designed to solve? Perhaps, but I doubt it. Since Frege's puzzle has been laid on top of the *de se* puzzle, getting the right result involves combining what is needed for both – Lewis's self-ascription plus, for example, Perry's, or Salmon's, ways of believing/ascribing.[36]

The next challenge is similar, except that the property self-ascribed in necessarily empty. In this case, adapted from (Perry 1979), Perry is shopping in the supermarket, when he looks up at two differently placed security mirrors and says to himself "I am nearer to him [demonstrating the man seen in mirror 1] than I am to him [demonstrating the man seen in mirror 2]." Since he is the man seen in both mirrors, he thereby self-ascribes *being nearer to Perry than to Perry*. Because this property is necessarily empty, self-ascription of it will, for Lewis, result in believing the impossible, which, for

him, is itself impossible.[37] Since this is wrong, salvaging the Lewisian *de se* requires embracing hyperintensionalism. Not surprising.

Next notice that John and Mary express *different beliefs* by sincerely uttering "I am hungry" – despite the fact that they self-ascribe the *same property*. How does this fit the idea that *de se* belief is the self-ascription of properties? Pretty well, if we add that an agent x who *self-ascribes P*, also counts as *ascribing* P of *x* (but not conversely). Whereas both involve predicating P of x, the former requires *thinking of the predication target in the first-person way* (whatever that amounts to), while the latter doesn't. It will then follow that, in addition to their identical *de se* beliefs, John and Mary also have different *de re* beliefs. (I will return to the question of whether they really do also express *identical* beliefs.)

The same idea allows us to recognize that Lingens and his friend Lola express the same (*de re*) belief when Lingens sincerely says, "this book is about me," and Lola, agrees, saying "That book is about you." A state of *de se* believing something is always also a state of *de re* believing something closely related. Innocent enough, this idea worsens the problems posed by Lewis's anti-hyperintensionalism – making it impossible for Lingens, prior his epiphany, to coherently take himself *not to be Lingens*. Since doing that would involve both self-ascribing *not being Lingens* and believing the necessary falsehood *that Lingens is not Lingens*, Lewis must, by the assumptions in the previous footnote, (wrongly) disallow the possibility that Lingens might wrongly believe *that he isn't Lingens*. To avoid this, hyperintensionality must (again) be embraced.

So far, I have written as if Lewis took the objects of *de se* beliefs to be properties and the objects of other beliefs to be propositions. He didn't. Rightly wanting a unified analysis, and judging propositions not to be objects of *de se* beliefs, he took properties to be the objects of all beliefs. On this view, to believe that the earth moves is to self-ascribe *being such that the earth moves* (a property everything has). This is artificial at best, and problematic at worst. For one thing, Lewisian properties aren't true or false; for another, agents who self-ascribe *being hungry* don't thereby express identical beliefs. To me, this suggests the need for a new conception of propositions that explains how genuine *de se* propositions (the entertaining of which require Lewisian self-ascription) can be distinct from, but representationally identical with, corresponding *de re* propositions. Elsewhere, I argue for such a conception, which preserves the virtues of Lewis's solution without its drawbacks.[38]

Another brilliant and highly influential paper is Lewis (1975b). Marrying Lewis (1970b, and 1969), it reconciles two ways of thinking about language. According to one, languages are sets of well-formed, interpreted, expressions; according to the other they are social practices. As Lewis persuasively argues, both perspectives are needed. By distinguishing the question *What is a language?* – an abstract object with semantic properties – from the question *What facts are necessary and sufficient for a population to speak it?*, he puts the two perspectives into a single frame. In the simplest case (ignoring ambiguity and indexicality, and considering only declarative sentences) he takes a language L to be a set of well-formed sentences paired with their meanings – which are sets of worlds at which they are true. (Meanings of constituents of sentences are not part of the specification because Lewis shared Quine's worries about referential indeterminacy.) For L to be used by a population P is for *a convention of truthfulness and trust in L* to exist in P. To be *truthful in L* is, to a first approximation, to assertively utter a sentence S of L only when one believes S to be *true in L*. To be *trusting in L* is, roughly, to be disposed to believe that others in P are *truthful in L*, and so will assertively utter S only if they believe S to be *true in L*. From this it follows that the trusting hearer will be disposed to believe that S is *true in L*, to the extent she has reason to believe that speaker knows what he is talking about. Finally, to say that there is a *convention* of *truthfulness and trust in L* is to say, very roughly, (i) that there is a regularity of such among members of P, (ii) that members of P realize this, (iii) that the expectation that others conform to the regularity give one a good reason to conform, (iv) that there is a general preference, given that most conform, for all (including oneself) to conform, and (v) that

members of P could, in theory, satisfy their interests in communication by conforming to an alternative regularity of truthfulness and trust in a different language.

These are, of course, just the bare bones of Lewis's theory, without his many complex elaborations. It is, however, enough to raise two questions.

Q1 Is it essential that something like his regularities be *conventions*?

Q2 Does the framework require individuating meanings/propositions so coarsely, or is it compatible with a genuinely hyperintensional conception of both?

First Q1. Conditions (iii) and (iv) are crucial to Lewis's analysis of convention and to his view that the solution to the coordination problem provided by *linguistic conventions* gives members of P good practical and epistemic *reasons* for action and belief. In what sense does a regularity of truthfulness and trust in L provide each speaker with such reasons? Let's grant that speaker x has an abiding interest in communication and in the conformity of others to the policy of being *truthful and trusting in L*. Presumably, the fact that they do conform gives x a good reason to trust that they do. Does it also give x a good reason to be *truthful in L*? Although it may give x a moral reason, this is not the type of reason Lewis emphasizes, any more than the fact that there is a moral reason – to avoid injuring or inconveniencing others – to conform (when driving in the USA) to the regularity of driving on the right-hand side of the road is the main sort of interest served by the convention of driving on the right. In the driving case, anyone interested in getting safely and swiftly from point A to point B has reason to conform, independent of a concern for others. Although the issue is delicate in the linguistic case, it would seem that Lewis needs a similarly compelling reason that comes from x's interest in communication. Is there such a reason for x to speak *truly in L*? Pending an answer to this question, it is not easy to say whether conditions (iii) and (iv) of Lewis's definition of a convention are met.[39]

The answer to Q2 is more direct. Substituting hyperintensional propositions for Lewis's intensional ones strengthens the analysis. Suppose, what I take to be obvious, that necessarily equivalent sentences may differ both in meaning and in the beliefs they are used to express. It follows naturally that L and L* may differ in the hyperintensional propositions they assign to the same sentences, even though the truth conditions of sentences resulting from those assignments are necessarily equivalent. This in turn requires an adjustment in the statement of the convention of truthfulness and trust. We now say (again very roughly): to be truthful in L is to assertively utter a sentence S of L only when one believes the proposition S expresses in L, and to be trusting in L is to be disposed to believe that other members of P are truthful in L too. With this modest reformulation, none of the attractiveness of Lewis's guiding idea is lost.

Lewis (1979c) is a similarly brilliant and even more influential discussion of the nuanced relationship between meaning, context, and assertion. The article develops two leading ideas. First, various linguistic items and constructions place requirements on contexts in which sentences containing them can acceptably be uttered – requirements regarding what must be assumed or presupposed by conversational participants prior to the utterance, what is taken by them to be under discussion, what possibilities are relevant for conversational exploration, the standards of precision governing the discussion of certain topics and the uses of certain words, the objects most saliently available for conversational reference, and the like. Second, when an utterance requires some aspect of the context to meet one of these requirements which is not, in fact, met prior to the utterance, there is a general, though defeasible, presumption that the conversational participants will *accommodate* the speaker by updating the context so as to bring it into line with what the speaker's remark requires (if doing so is unobjectionable and doesn't involve violating their other commitments). Lewis applies this idea

with considerable effect to presuppositions, uses of singular definite descriptions, uses of pronouns seemingly (but not really) anaphoric on prior uses of quantifiers, the reference points of narratives, uses of vague predicates, the choice of different and varying modal domains for terms like 'can' and 'must', the choice of relevant alternatives that must be ruled out to verify claims to know a proposition to be true, the use of language in developing complex plans, and the use of explicit performatives that are verified (made true by) that very use.[40] The influence of this article in epistemology, philosophy of language, and linguistic semantics can hardly be overestimated.

Lewis was, of course, the author of many more influential articles in the philosophy of language. A final one that must be mentioned is Lewis (1975a). In it, he argues that sentence adverbs such as 'always', 'usually', 'often', sometimes', and 'never' are unselective quantifiers that bind all free variables in their scope. For example, in 'Usually (x likes y iff y likes x)' 'usually' binds occurrences of both 'x' and 'y', so the sentence is true iff for most pairs of people the first likes the second iff the second likes the first. In addition to being attractive, this analysis has also been extremely influential in the literature in linguistic semantics. For example, as noted in Richard Holton (2003), it is cited in Irene Heim (1990) as providing important inspiration for discourse representation theory.

In sum, despite what I judge to be the less than successful outcome of David's overarching metaphysical project, there can be no scanting his many large and lasting contributions to a variety of areas, including, but not restricted to, philosophical logic, the philosophy of language, and linguistic semantics. Nor was his influence limited to the dazzling corpus of his published work. For 30 years he was, during the era of its greatness, a pillar of Princeton philosophy department. A cooperative and influential colleague, as well as a dedicated and conscientious teacher, his impact on the PhD program there was profound. Always one of the chief draws in recruiting graduate students and one of the strongest influences on their education, he was, for nine years, a model director of graduate studies. His dissertation students are spread far and wide across the profession, and his personal influence on individual philosophers in the USA, "down under," and throughout the English-speaking world (and beyond) was simply unmatched during his far too short a time on Earth.

Acknowledgments

Thanks to Ali Kazmi for very helpful comments on this paper.

Notes

1 Though they did so in very different ways. Whereas Lewis continued the Quinean identification of necessity with analyticity, Kripke did not. In Lewis (1969, 1970, 1975b) analyticity is necessity, which, in light of his "modal realism" is explained in terms of ordinary quantification *over everything that is*. By contrast, for Kripke necessity is primitive but linked to the traditional metaphysical doctrine of essentialism.

2 See Soames 2007a, chapter 5 of 2010a, and 2010b.

3 On pp. 5–6 of Lewis (1986a) Lewis says that the operator *in w* is like the operator *in Australia*: "In w, HH is having dinner" is like "In Australia, Hubert is having dinner." Just as the latter says that Hubert is located in Australia and is having dinner, so the former says that HH inhabits w and is having dinner.

4 See Lewis 1986a, 199–201, where he argues in this way that the idea that an individual might have different intrinsic properties at different worlds is unintelligible.

5 Lewis 1986a, 196.

6 (i) is counterpart theory, if *being so-and-so* is a "pure" property that doesn't itself involve any individuals; if it does involve an individual o, then the property had by the Humphrey counterpart at w will involve an o counterpart, and so will not be identical with *being so-and-so*.

7 Lewis 1986a, 197.
8 See Gideon Rosen 1990, 1995.
9 Quine 1960, 220–1.
10 Although Quine was willing to accept physical theories at face value, he never explicitly repudiated holistic verificationism, and was, it would appear, never fully the scientific realist that Lewis was.
11 See lectures 2 and 3 of Kripke 1980.
12 For one such version of the Kripke story see Soames 2007c.
13 See, for example, Lewis 1966, 1970c, 1974, 1997.
14 If one is worried that some counterfactuals with metaphysically impossible antecedents should come out false, while others are really true, the Kripkean may go beyond Lewis by including world-states that are epistemically, but not metaphysically, possible in stating the truth conditionals of counterfactuals. More on such world-states below.
15 See Kit Fine 1975.
16 See the discussion in Lewis 1986b, e.g. at p. ix.
17 Soames 2006a.
18 To the extent that Kripkean intuitions about rigid designation need to be accommodated, Lewis came (by the 1990s) to see rigidified descriptions plus Chalmers/Jackson style two-dimensionalist analyses as the way to do so. The appeal of this strategy was, as explained in Soames (2005, 2006a), to preserve a pre-Kripkean conception of conceptual possibility as the only possibility. For Lewis (and the others) the distinction between epistemic and metaphysical modalities which allows a single proposition to be conceivable though necessarily false is replaced by a system in which a pair of propositions (associated with the same sentence) can receive different modal evaluations.
19 Page 122 of the version reprinted in Lewis 1983a.
20 Lewis 1983a, 123.
21 For discussion of the rigidity of general terms see Soames 2007c.
22 It is important to distinguish the question whether 'pain' is a rigid designator from the question whether being a pain is an essential property of anything that has it. Although Kripke seems to run them together in his discussion of pain in *Naming and Necessity* (1980, 148–9), they are distinct (see Soames 2006, 181–2, and 2002, 252–3). Having dealt with the rigidity issue, we may wonder whether, as Kripke thinks, *being a pain* is an *essential property* of anything that has it. Consider, the intense headache I had last night. Could it have existed without being a pain? It's mildly bizarre to suppose it could – while being pleasurable or unfelt. Lewis seems to allow both. Suppose at w, I exist with all my C-fibers, but my brain has been altered by the addition of genetically designed D-fibers that make the C-fiber stimulation pleasurable. At w, the same C-fibers fire that constituted my headache at @, while being experienced as a pleasurable tickling. It seems, on Lewis's account, that the thing which is my headache at @ also exists at w. Is it a pain? If others of my kind haven't had the benefit of the procedure, Lewis's theory will count me as feeling mad pain, even though it is enjoyable. That seems wrong. Suppose I am the only one of my kind at w, or that all the others have new D-fibers too. In both cases, my headache, which for Lewis is identical with a certain firing of C-fibers, exists and is *not a pain*, but rather is a pleasurable sensation. Not, I think, a convincing counterexample to Kripke's essentiality thesis. By contrast, the example causes no problem for our final, purely functionalist reformulation of Lewis, since on this analysis the pain-state is *realized* by, but not *identified* with, the C-fiber firing.
23 Russell's system is examined in detail in chapter 12 of volume 1 of Soames 2014. I don't put the metaphysical system of the *Tractatus* in the same category because of its subservience to the larger themes of the work, which were devoted to turning philosophy away from genuine metaphysics.
24 In addition to establishing this result, the article also offers an analysis of indicative conditionals on which [If A, then B] has the truth conditions of a material conditional, while having assertability conditions requiring the conditional probability B given A to be high. In the postscript (Lewis 1986c), he gives up the details of this view, in favor of a related account given in Frank Jackson 1979, 1981.
25 The demonstration, in Hans Kamp (1971), that times and worlds must be double-indexed had not been widely recognized. David Kaplan's reconceptualization (including his distinction between character and content) was also not yet in place. The indices in Lewis (1970b) each consisted of a single time, place, world,

and assignment of objects to variables. In Lewis (1980b, 1983b) he noted the work of Kamp, Stalnaker, and Kaplan, and recognized various ways in which his original parameters might be enriched.

26 See Soames 1987, 2008.

27 See Soames 2006b and chapter 3 of King et al. 2014.

28 In Lewis 1997, we are given the following account of the descriptions to be rigidified in the case of names.

> Did not Kripke and his allies refute the description theory of reference, at least for names of people and places? . . . I disagree. What was well and truly refuted was a version of descriptivism in which the descriptive senses were supposed to be a matter of famous deeds and other distinctive peculiarities. A better version survives the attack: *causal descriptivism*. The descriptive sense associated with a name might for instance be 'the place I have heard of under the name "Toromeo" ' or maybe 'the causal source of this token: Toromeo', and for the account of the relation being invoked here, just consult the writings of causal theorists of reference. (Lewis 1999, 352, fn. 22)

The final clause of this passage is problematic. There is no precise theory of how the referent of one's use of a name is inherited from other uses. Since it is inherited, the description "the referent of the uses from which my use of 'Toromeo' somehow inherits its reference" picks out the right object, but can't, on pain of circularity, be what *determines* my referent. The problem for causal descriptivism is that speakers have no other, non-circular, descriptive means of fixing reference. For more on this and related problems with attempts to fix the referents of names descriptively see my reply to Jackson in Soames 2007b.

29 See Soames 2005.

30 Throughout the 1999–2000 academic year David and I jointly taught a graduate seminar on *Naming and Necessity* and its aftermath. In it, he identified propositions with sets of worlds, and used diagonalization plus a Jackson/Chalmers-like distinction between primary and secondary intension to deal with attitude ascriptions.

31 The function that assigns "worlds" 1–3 truth, and all others falsity, isn't intentional either. What, after all, are truth and falsity but properties we grasp primarily through their application to propositions? But surely, if propositions are needed to illuminate truth and falsity, they can't be among the building blocks for constructing propositions. See chapters 3 and 6 of King *et al.* 2014.

32 See Soames 2006a.

33 See Soames 2011.

34 The semantics of 'actual' is discussed in Lewis 1970. Although the analysis there is called "indexical," it doesn't take into account the then unrecognized need for double indexing of times and worlds. Because of this, only one of the two meanings he assigns to 'is actual' (and to 'the actual world') is genuinely indexical – the reading on which 'is actual' is true of a world iff it is identical with the world of the context (assuming double indexing). The other, non-indexical, reading is one in which 'is actual' is true (relative to a context and a circumstance of evaluation) of a world w iff w is the world of the circumstance. The obvious truths that Lewis is forced to deny are those involving the genuine indexical reading. He is forced to deny them by the basic tenet of modal realism that for Hubert to possibly exist, or be so-and-so, is for an inhabitant HH, of a merely possible world, to exist, or be so-and so. That is the key mistake; adopting an approximately correct indexical semantics for 'actual' merely helps reveal it.

35 See Soames 2007a and 2010, chapter 6.

36 Salmon 1984.

37 For Lewis, when S is necessarily false, ⌜A believes that S⌝ is necessarily equivalent to ⌜A believes that S & P⌝ for every P. Since 'believe' distributes over conjunction, the truth of the former guarantees the truth of ⌜A believes that P⌝ *for every P*. Since it is impossible for any agent to simultaneously believe everything (and its negation), it must be impossible for ⌜A believes that S⌝ to be true when S is necessarily false.

38 See chapter 6 of King *et al.* 2014.

39 See Lewis (1983, 184–5) for his interesting and nuanced discussion of this issue.

40 Lewis acknowledges Stalnaker (1973, 1974) on presupposition, Kratzer (1977) on relative modality, and Lemmon (1962) on performatives. According to Lemmon and Lewis, declarative sentences like 'I promise to arrive on time' and 'I order you to leave' are used to express propositions that are made true by asserting them (in the right circumstances). To assert them in such circumstances is to perform the speech acts of

promising and ordering. In addition to further elaborating these and other analyses, Lewis is responsible for bringing them together under the common head of *accommodation*.

References

Barwise, J. & Perry, J. (1983). *Situations and attitudes*. Cambridge, MA: MIT Press.

Bennett, J. (2003). *A philosophical guide to conditionals*. Oxford: Clarendon Press.

Carnap, R. (1928). *Der Logische Aufbau der Welt*. Leipzig: Felix Meiner Verlag. Trans. R.A. George (1967), *The logical structure of the world, pseudoproblems in philosophy*, Berkeley and Los Angeles: University of California Press.

Carnap, R. (1950). Empiricism, semantics, and ontology. *Revue Internationale de Philosophie*, 4, 20–40. Revised and reprinted in *Meaning and necessity*, 2nd ed. (1956). Chicago: University of Chicago Press.

Castaneda, H.-N. (1966). He: A study in the logic of self-consciousness. *Ratio*, 8, 130–57.

Castaneda, H.-N. (1968). On the logic of attributions of self-knowledge to others. *Journal of Philosophy*, 65, 439–56.

Chalmers, D. (1996). *The conscious mind*. New York: Oxford University Press.

Edgington, D. (2004). Counterfactuals and the benefit of hindsight. In P. Dowe & P. Noordhof (Eds.), *Causation and counterfactuals*, pp. 12–27. London: Routledge.

Fine, K. (1975). Critical notice of *Counterfactuals*. *Mind*, 84, 451–8.

Heim, I. (1990). E-type pronouns and donkey anaphora. *Linguistics and Philosophy*, 13, 137–77.

Holton, R. (2003). David Lewis's philosophy of language. *Mind and Language*, 18, 286–95.

Jackson, F. (1979). On assertion and indicative conditionals. *Philosophical Review*, 88, 565–89.

Jackson, F. (1981). Conditionals and possibilia. *Proceedings of the Aristotelian Society*, 81, 125–37.

Jackson, F. (1998). *From metaphysics to ethics*. Oxford: Oxford University Press.

Kamp, H. (1971). Formal properties of 'now'. *Theoria*, 37, 227–74.

King, J., Soames, S., & Speaks, J. (2014). *New thinking about propositions*. Oxford: Oxford University Press.

Kment, B. (2006). Counterfactuals and explanation. *Mind*, 115, 261–310.

Kratzer, A. (1977). What 'must' and 'can' must and can mean. *Linguistics and Philosophy*, 1, 337–55.

Kripke, S. (1980). *Naming and necessity*. Cambridge, MA: Harvard University Press.

Lemmon, E.J. (1962). On sentences verifiable by their use. *Analysis*, 22, 86–9.

Lewis, D. (1966). An argument for the identity theory. *Journal of Philosophy*, 63, 17–25. Reprinted in Lewis (1983a).

Lewis, D. (1969). *Convention*. Cambridge, MA: Harvard University Press.

Lewis, D. (1970a). Anselm and actuality. *Nous*, 4, 175–88. Reprinted in Lewis (1983a).

Lewis, D. (1970b). General semantics. *Synthese*, 27, 18–67. Reprinted in Lewis (1983a).

Lewis, D. (1970c). How to define theoretical terms. *Journal of Philosophy*, 67, 427–46. Reprinted in Lewis (1983a).

Lewis, D. (1973a). *Counterfactuals*. Cambridge, MA: Harvard University Press.

Lewis, D. (1973b). Causation. *Journal of Philosophy*, 70, 556–67. Reprinted in Lewis (1986b).

Lewis, D. (1974). Radical interpretation. *Synthese*, 23, 331–44. Reprinted in Lewis (1983a).

Lewis, D. (1975a). Adverbs of quantification. In E.L. Keenan (Ed.), *Formal semantics of natural language*, pp. 3–15. Cambridge: Cambridge University Press. Reprinted in Lewis (1998).

Lewis, D. (1975b). Languages and language. In K. Gunderson (Ed.), *Language, mind, and knowledge*, pp. 3–35. Minneapolis: University of Minnesota Press. Reprinted in Lewis (1983a).

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *The Philosophical Review*, 85, 297–315. Reprinted in Lewis (1986b).

Lewis, D. (1979a). Attitudes *de dicto* and *de se*. *The Philosophical Review*, 88, 513–43. Reprinted in Lewis (1983).

Lewis, D. (1979b). Counterfactual dependence and time's arrow. *Nous*, 13, 418–46. Reprinted in Lewis (1986b).

Lewis, D. (1979c). Scorekeeping in a language game. *Journal of Philosophical Logic*, 8, 339–59.

Lewis, D. (1980a). Mad pain and Martian pain. In N. Block (Ed.), *Readings in the philosophy of psychology*, Vol. 1, pp. 216–22. Cambridge, MA: Harvard University Press. Reprinted in Lewis (1983a).

Lewis, D. (1980b). Index, context, and content. In S. Kanger and S. Öhman (Eds.), *Philosophy and grammar*, pp. 79–100. Dordrecht: Reidel.
Lewis, D. (1983a). *Philosophical papers*, Vol. I. New York: Oxford University Press.
Lewis, D. (1983b). Postscript to "General semantics." In Lewis (1983a).
Lewis, D. (1986a). *On the plurality of worlds*. Oxford: Blackwell.
Lewis, D. (1986b). *Philosophical papers*, Vol. II. New York: Oxford University Press.
Lewis, D. (1986c). Postscript to "Probabilities of conditionals and conditional probabilities." In Lewis (1986b).
Lewis, D. (1996). Elusive knowledge. *Australasian Journal of Philosophy*, 74, 549–67. Reprinted in Lewis (1999).
Lewis, D. (1997). Naming the colors. *Australasian Journal of Philosophy*, 75, 325–42. Reprinted in Lewis (1999).
Lewis, D. (1998). *Papers in philosophical logic*. Cambridge: Cambridge University Press.
Lewis, D. (1999). *Papers in metaphysics and epistemology*. Cambridge: Cambridge University Press.
Lewis, D. (2004). Causation as influence. In J. Collins, N. Hall, and L.A. Paul (Eds.), *Causation and counterfactuals*, pp. 75–106. Cambridge, MA: MIT Press.
Perry, J. (1977). Frege on demonstratives. *Philosophical Review*, 86, 474–97.
Perry, J. (1979). The problem of the essential indexical. *Noûs*, 13, 3–21.
Quine, W.V.O. (1948). On what there is. *Review of Metaphysics*, 2, 21–38. Reprinted in Quine (1980).
Quine, W.V.O. (1951). Two dogmas of empiricism. *Philosophical Review*, 60, 20–43. Reprinted in Quine (1980).
Quine, W.V.O. (1960). *Word and object*. Cambridge, MA: MIT Press.
Quine, W.V.O. (1969). Reply to Chomsky. In D. Davidson and J. Hintikka (Eds.), *Words and objections*, p. 303. Dordrecht: Reidel.
Quine, W.V.O. (1970). On the reasons for indeterminacy of translation. *Journal of Philosophy*, 67, 178–183.
Quine, W.V.O. (1980). *From a logical point of view*, 2nd ed. Cambridge, MA: Harvard University Press.
Rosen, G. (1990). Modal fictionalism. *Mind*, 99, 327–54.
Rosen, G. (1995). Modal fictionalism fixed. *Analysis*, 55, 67–73.
Russell, B. (1914). *Our knowledge of the external world*. Chicago and London: Open Court.
Russell, B. (1918–19). The philosophy of logical atomism. *Monist*, 5(28), 495–527, continued in *Monist*, 5(29), 32–63, 190–222, 345–80. Reprinted in B. Russell (1985) *The philosophy of logical atomism*. Peru, IL: Open Court.
Salmon, N. (1984). *Frege's puzzle*. Cambridge, MA: MIT Press.
Schaffer, J. (2004). Counterfactuals, causal independence and conceptual circularity. *Analysis*, 64, 299–309.
Soames, S. (1987). Direct reference, propositional attitudes, and semantic content. *Philosophical Topics*, 15, 44–87. Reprinted in Soames (2009).
Soames, S. (2002). *Beyond rigidity*. New York: Oxford University Press.
Soames, S. (2005). *Reference and description*, Princeton, NJ: Princeton University Press.
Soames, S. (2006a). The philosophical significance of the Kripkean necessary aposteriori. *Philosophical Issues*, 16, 288–309. Reprinted in Soames (2009).
Soames, S. (2006b). Understanding assertion. In J.J. Thompson & A. Byrne (Eds.), *Content and modality: themes from the philosophy of Robert Stalnaker*, pp. 222–50. New York: Oxford University Press. Reprinted in Soames (2009).
Soames, S. (2007a). Actually. *Aristotelian Society Supplementary Volume*, 81, 251–77. Reprinted in Soames (2009).
Soames, S. (2007b). The substance and significance of the dispute over two-dimensionalism. *Philosophical Books*, 48, 34–49.
Soames, S. (2007c). What are natural kinds. *Philosophical Topics*, 35, 329–42.
Soames, S. (2008). Why propositions can't be sets of truth-supporting circumstances. *Journal of Philosophical Logic*, 37, 267–76. Reprinted in Soames (2009).
Soames, S. (2009). *Philosophical essays*, Vol. 2. Princeton, NJ: Princeton University Press.
Soames, S. (2010a). *Philosophy of language*. Princeton, NJ: Princeton University Press.
Soames, S. (2010b). True at. *Analysis*, 71, 124–33.
Soames, S. (2011). Kripke on metaphysical and epistemic possibility: two routes to the necessary aposteriori. In A. Berger (Ed.), *Saul Kripke*, pp. 78–99. Cambridge: Cambridge University Press.
Soames, S. (2014). *The analytic tradition*, Vol. 1. Princeton, NJ: Princeton University Press.

Stalnaker, R. (1968). A theory of conditionals. In *Studies in Logical Theory*, American Philosophical Quarterly Monograph Series, No. 2, Oxford: Blackwell. Reprinted in W.L. Harper, R. Stalnaker, and G. Pearce (1981) *Ifs*. Dordrecht: Reidel.

Stalnaker, R. (1973). Presuppositions. *Journal of Philosophical Logic*, 2, 447–57.

Stalnaker, R. (1974). Pragmatic presuppositions. In M.K. Munitz & P.K. Unger (Eds.), *Semantics and Philosophy*, pp. 197–213. New York: New York University Press.

Stalnaker, R. (1978). Assertion. In P. Cole (Ed.), *Syntax and Semantics*, Vol. 9. New York: Academic Press. Reprinted in R. Stalnaker (1999) *Context and content: essays on intentionality in speech and thought*, pp. 78–95. New York: Oxford University Press.

Van Fraassen, B.C. (1980). *The scientific image*. Oxford: Oxford University Press.

Part III

Metaphysics and Science

8

Humean Supervenience

BRIAN WEATHERSON

8.1 What Is Humean Supervenience?

As with many aspects of David Lewis's work, it is hard to provide a better summary of his views than he provided himself. So the following introduction to what the Humean supervenience view is will follow the opening pages of Lewis (1994a) extremely closely. But for those readers who haven't read that paper, here's the nickel version.

Humean supervenience is the conjunction of three theses.

1. *Truth supervenes on being* (Bigelow 1988). That is, all the facts about a world supervene on facts about which individuals instantiate which fundamental properties and relations.
2. *Anti-haecceitism.* All the facts about a world supervene on the distribution of qualitative properties and relations; rearranging which properties hang on which "hooks" doesn't change any facts.
3. *Spatiotemporalism.* The only fundamental relations that are actually instantiated are spatiotemporal, and all fundamental properties are properties of points or point-sized occupants of points.

The first clause is a core part of Lewis's metaphysics. It is part of what it is for some properties and relations to be fundamental that they characterize the world. Indeed, Lewis thinks something stronger, namely that the fundamental properties and relations characterize the world *without redundancy* (Lewis 1986a, 60). This probably isn't true, for a reason noted in Sider (1993). Consider the relations *earlier than* and *later than*. If these are both fundamental, then there is some redundancy in the characterization of the world in terms of fundamental properties and relations. But there is no reason to believe that one is fundamental and the other isn't. And it is hard to see how we could give a complete characterization of the world without either of these relations. So we'll drop the claim that the fundamental properties relations characterize the world without redundancy, and stick to the weaker claim, namely that the fundamental properties and relations characterize the world completely.

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

The second clause is related to Lewis's counterpart theory. Consider what it would be like for anti-haecceitism to fail. There would have to be two worlds, with the same distribution of qualitative properties, but with different facts obtaining in each. These facts would have to be non-qualitative facts, presumably facts about which individual plays which role. So, perhaps, to use a well-known example, there could be a world in which everything is qualitatively as it is in this world, but in which Barack Obama plays the Julius Caesar role, and vice versa. So Obama conquers Gaul and crosses the Rubicon, Caesar is born in Hawai'i and becomes president of the United States. But what could make it the case that the Gaul-conqueror in that world is really *Obama*'s counterpart, and not Caesar's? Nothing qualitative, and nothing else it seems is available. So this pseudo-possibility is not really a possibility. And so on for all other counterexamples to anti-haecceitism.

The third clause is the most striking. It says there are no fundamental relations beyond the spatio-temporal, or fundamental properties of extended objects. If we assume that "properties" of objects with parts are really relations between the parts, and anything extended has proper parts, then the second clause reduces to the first. I think it isn't unfair to read Lewis as holding both those theses.

Since for Lewis the fundamental qualities are all intrinsic, the upshot is that the world is characterized by a spatiotemporal distribution of intrinsic qualities. As Lewis acknowledged, this was considerably more plausible given older views about the nature of physics than it is now. We'll return to this point at great length below. But for now the key point is to see the kind of picture Humean supervenience offers. The world is like a giant video monitor. The facts about a monitor's appearance supervene, plausibly, on intrinsic qualities of the pixels, plus facts about the spatial arrangement of the pixels. The world is four-dimensional, not two-dimensional like the monitor, but the underlying picture is the same.

8.2 Supervenience

Given the name Humean supervenience you might expect it to be possible to state Humean supervenience as a supervenience thesis. But this turns out to be hard to do. Here is one attempt at stating Humean supervenience as a supervenience thesis that is happily clear, and unhappily false.

> **Strong modal Humean supervenience** For any two worlds where the spatio- temporal distribution of fundamental qualities is the same, the contingent facts are the same.

But Humean supervenience does not make a claim this strong. It is consistent with Humean supervenience that there could be fundamental non-spatio-temporal relations. The only thing Humean supervenience claims is that no such relations are instantiated. In a pair of possible worlds where there are such relations, and the relations vary but the arrangement of qualities is the same, strong modal Humean supervenience will fail. In the Introduction to Lewis (1986b), he suggested the following weaker version.

> **Local modal Humean supervenience** For any two worlds at which no alien properties or relations are instantiated, if the spatiotemporal distribution of fundamental qualities is the same at each world, the contingent facts are also the same.

An alien property(/relation) is a fundamental property(/relation) that is not actually instantiated. So this version of Humean supervenience says that, to get a difference between two worlds, you have to either have a change in the spatio-temporal arrangement of qualities, or the instantiation of actually uninstantiated fundamental properties or relations.

But Lewis eventually decided that wouldn't do either. In response to Haslanger (1994), he conceded that enduring objects would generate counterexamples to local modal Humean supervenience even if there were no alien properties or relations. So he fell back to the following, somewhat vaguely stated, thesis. (See Lewis (1994a) for the concession, and Hall (2010) for an argument that he should not have conceded this to Haslanger.)

> **Familiar modal Humean supervenience** For any two "worlds like ours" (Lewis 1994a, 475), if the spatiotemporal distribution of fundamental qualities is the same at each world, the contingent facts are also the same.

What's a "world like ours"? It isn't, I fear, entirely clear. But this doesn't matter for the precise statement of Humean supervenience. The three theses are clear enough, and state what Humean supervenience is. The only difficulty is in stating it as a *supervenience* thesis.

8.3 What Is Perfect Naturalness?

That definition does, however, require that we understand what it is for some properties and relations to be *fundamental*, or, as Lewis put it following his discussion in Lewis (1983), *perfectly natural*. The perfectly natural properties and relations play a number of interconnected roles in Lewis's metaphysics and his broader philosophy.

Most generally, they characterize the difference between real change and "Cambridge change," and the related difference between real similarity, and mere sharing of grue-like attributes. This somewhat loose idea is turned, in *Plurality*, into a definition of duplication.

> . . . two things are duplicates iff (1) they have exactly the same perfectly natural properties, and (2) their parts can be put into correspondence in such a way that corresponding parts have exactly the same perfectly natural properties, and stand in the same perfectly natural relations. (Lewis 1986a, 61)

The intrinsic properties are then defined as those that are shared between any two (possible) duplicates. So, as noted above, Humean supervenience says that the spatiotemporal distribution of intrinsic features of points characterizes worlds like ours.

I've gone back and forth between describing these properties as fundamental and describing them as perfectly natural. And that's because for Lewis, the perfectly natural properties are in a key sense fundamental. For reasons to do with the nature of vectorial properties, I think this is probably wrong (Weatherson 2006). That is, we need to hold that some derivative properties are perfectly natural in order to get the definition of intrinsicness in terms of perfect naturalness to work. But for Lewis, the perfectly natural properties and relations are all fundamental.

Part of what Lewis means by saying that some properties are fundamental is that all the facts about the world supervene on the distribution. (This is Bigelow's thesis that truth supervenes on being.) But I think he also means something stronger. The non-fundamental facts don't merely supervene on the fundamental facts; those non-fundamental facts are true *because* the fundamental facts are true, and *in virtue of* the truth of the fundamental facts.

The perfectly natural properties play many other roles in Lewis's philosophy besides these two. They play a key role in the theory of laws, for instance. They are a key part of Lewis's solution to the New Riddle of Induction (Goodman 1955). And they play an important role in Lewis's theory of content, though just exactly what that role is is a matter of some dispute. (See Sider (2001) and Weatherson (2003) for one interpretation, and Schwarz (2009) for a conflicting interpretation.)

Now it is a pretty open question whether any one division of properties can do all these roles. One way to solve the New Riddle (arguably Lewis's way, though this is a delicate question of interpretation) is to be a dogmatist (in the sense of Pryor 2000) about inductive projections involving a privileged class of properties. Lewis's discussion of the New Riddle at the end of Lewis (1983) sounds like he endorses this view, with the privileged class being the very same class as fundamentally determines the structure of the world, and makes for objective similarity and difference. But why should these classes be the same? It might make more sense to, for instance, endorse dogmatism about inductive projections of *observational* properties, rather than about microphysical properties.

Lewis doesn't attempt to give a theoretically neutral definition of the perfectly natural properties. Rather, the notion of a perfectly natural property is introduced by the theoretical role it serves. But that theoretical role is very ambitious, covering many areas in metaphysics, epistemology, and the theory of content. We might wonder whether claims like Humean supervenience have any content if it turns out nothing quite plays that theoretical role. I think there is still a clear thesis we can extract, relying on the connection between intrinsicness and naturalness. It consists of the following claims:

- There is a small class of properties and relations such that the contingent facts at any world supervene on the distribution of these properties and relations.
- Each of these properties is an intrinsic property.
- At the actual world, the only relations among these which are instantiated are spatiotemporal, and all the contingent facts supervene not merely on the distribution of fundamental qualities and relations, but also on the distribution of fundamental qualities and relations *over points and point-sized occupants of points*.

Those theses are distinctively Lewisian, they are clearly entailed by Humean supervenience as Lewis conceives of it, they are opposed in one way or another by those who take themselves to reject Humean supervenience, but they are free of any commitment to there being a single class of properties and relations that plays all the roles Lewis wants the perfectly natural properties and relations to play. So from now on, when I discuss the viability of Humean supervenience, I'll be discussing the viability of this package of views.

8.4 Humean Supervenience and Other Humean Theses

Lewis endorsed many views that we might broadly describe as "Humean." Of particular interest here are the following three.

- *Humean supervenience*.
- *Nomological reductionism*. Nomological properties and relations (including lawhood, chance and causation) are not among the fundamental properties and relations.
- *Modal combinatorialism*. Roughly, anything can coexist with anything else.

We've stated modal combinatorialism extremely roughly, and will persist with using a fairly informal version of it throughout. For an excellent study of more careful versions of it, see Nolan (1996). But those details aren't as important to this debate. What is important for now is that all three of these theses are associated with what are known as Humean approaches to metaphysics in the contemporary literature. But how closely connected are they to each other, or for that matter to Hume?

One question about Humean supervenience is just how it connects to the work of the historical Hume. This would be a little easier to answer if there was a broad scholarly consensus that Hume actually believed the kind of simple regularity thesis of causation that Lewis attributes to him at the start of Lewis (1973). But it isn't clear that this is Hume's view (Strawson 2000). What is true is that Hume was skeptical that we could know more about causation than that it was manifested in certain distinctive kinds of correlations. But it is a further step to say that Hume inferred that causation just consists of these distinctive kinds of correlations.

A second question is how Humean supervenience, which perhaps should be referred to as so-called "Humean supervenience," or, perhaps even better, as "Lewisian supervenience," relates to the kind of regularity theory that Lewis attributes to Hume, or to the prohibition on necessary connections between distinct existences that underlies modal combinatorialism. Lewis seemed to see the three theses as related. Here he is explaining how he chose to name Humean supervenience (and recall that this *isn't* backed up by any detailed exegesis of Hume).

> Humean Supervenience is named in honour of the great denier of necessary connections. It is the doctrine that all there is to the world is a vast mosaic of local matters of particular fact just one little thing and then another. (Lewis 1986b, ix)

This is a slightly confusing passage, since it isn't clear why a violation of Humean supervenience would constitute a necessary connection of any kind. We will return to this point below. But it does seem to make clear that Lewis thought that Humean supervenience and modal combinatorialism were connected, since modal combinatorialism is much more closely connected to the denial that they can be necessary connections between distinct existences.

Compare how Lewis introduces Humean supervenience when discussing the role of possible worlds in formulating trans-world supervenience theses in *Plurality*.

> Are the laws, chances, and causal relationships nothing but patterns which supervene on this point-by-point distribution of properties? Could two worlds differ in their wars without differing, somehow, somewhere, in local qualitative character? (I discuss this question of "Humean Supervenience", inconclusively, in the Introduction to my Philosophical Papers, volume II.) (Lewis 1986a, 14)

This seems to connect Humean supervenience closely to nomological reductionism, since it makes the reducibility of the nomological properties and relations central to the question of whether Humean supervenience is true. We can also, I think, see Lewis connecting modal combinatorialism and nomological reductionism in a later passage in *Plurality* where he discusses why he doesn't believe that laws are necessary truths.

> Another use of [modal combinatorialism] is to settle – or as opponents might say, to beg – the question whether the laws of nature are strictly necessary. They are not…Episodes of bread-eating are possible because actual; as are episodes of starvation. Juxtaposed duplicates of the two, on the grounds that anything can follow anything; here is a possible world to violate the law bread nourishes.…It is no surprise that [modal combinatorialism] prohibited strictly necessary connections between distinct existences. What I have done is to take a Humean view about laws and causation, and use it instead as a thesis about possibility. Same thesis, different emphasis. (Lewis 1986a, 91)

So, for Lewis, these three theses are meant to be closely connected. And it is true that in the contemporary literature all three of them are frequently described as "Humean" theses. (Or at least they are so described in metaphysics and philosophy of science; again, we're bracketing questions of historical interpretation here.) But, on second glance, it isn't as clear what the connection between the three

theses could amount to. One immediate puzzle is that Humean supervenience is for Lewis a *contingent* thesis, while the other two theses are necessary truths. The accounts of causation, lawhood, and chance that he gives in defending nomological reductionism are clearly meant to hold in all kinds of worlds, not just worlds like ours. (Consider the amount of effort that is spent in Lewis (2004a) on defending the theory of causation from examples involving wizards, action at a distance, and so on.) And the formulation of modal combinatorialism in *Plurality* leaves little doubt that it is meant to be necessarily true.

This difference in modal status means that the theses can't be in any way equivalent. But you might think that they are in some way reinforcing. Even that isn't so clear. Consider the most dedicated kind of denier of modal combinatorialism, namely the fatalist who thinks that every truth is a necessary truth. She will *endorse* Humean supervenience. After all, she thinks that all the truths about the world supervene on any category of truths whatsoever, so they'll supervene on intrinsic properties of point-sized objects.

In the other direction, failures of Humean supervenience don't motivate compromising modal combinatorialism. Imagine a world where occasionally there are pairs of people who can know what each other is thinking, even though there is no independent informational chain between the two of them. It is just that a telepathic connection exists. Moreover, there is no rhyme or reason to when a pair of people will be telepathic; it is simply the case that some pairs of people are. In such a world, it is plausible that *being a telepathic pair* will be a fundamental relation. That's not a problem for Humean supervenience, since there aren't any such pairs in this world. But it does mean Humean supervenience is false in that world.

Assume that Daniels and O'Leary are a telepathic pair. Any duplication of the pair of them will also be telepathic, since, by Lewis's preferred definition of duplication, duplication preserves all fundamental properties and relations. Does that mean there's a necessary connection between Daniels and O'Leary? Not really. The spirit of modal combinatorialism is that you can duplicate any parts of any worlds, and combine them. One part of our world is Daniels. A duplicate of him need not include any telepathic connection to O'Leary; indeed, he has duplicates in worlds in which O'Leary is absent. Another part of the world is O'Leary; duplicates of him need not include a connection to Daniels. Putting the two together, there is a world where there are duplicates of Daniels and O'Leary, but no telepathic connection between the two. So modal combinatorialism suggests that, even when Humean supervenience fails, there won't be a necessary connection between distinct objects. So Humean supervenience really isn't that important to the idea that there are no necessary connection between distinct existences.

What's closer to the truth, I think, is that Humean supervenience is *interesting* because of modal combinatorialism. If modal combinatorialism fails, then Humean supervenience doesn't capture anything important. In particular, it doesn't capture the idea that the nomic is somehow less fundamental than (some features of) the non-nomic. It is only given modal combinatorialism that we can make these kinds of priority claims in modal terms. Think about the philosopher who denies modal combinatorialism on the grounds that laws of nature are necessarily true. That philosopher will say that the laws supervene on the distribution of intrinsic properties of points, because the laws supervene on any set of facts that you like. But they will deny that this makes the distribution of intrinsic properties of points more fundamental than the laws. It is only given modal combinatorialism that we can claim that supervenience theses are any guide whatsoever to fundamentality.

What about the connection between nomological reductionism and Humean supervenience? It can't be equivalence, since Lewis agrees that Humean supervenience fails in worlds in which nomological reductionism is true. For the same reason, it can't be that failures of Humean supervenience entail failures of nomological reductionism. What about the other direction? Could we imagine nomological reductionism failing while Humean supervenience holds? I think this is a coherent

possibility, but not at all an attractive one. (Compare, in this respect, the discussion of theories that "qualify technically as Humean" at Lewis 1994a, 485.) It requires that some of the irreducible, nomological properties be intrinsic properties of point-sized objects. Well, we could imagine two worlds where *F* and *G* are coextensive, intrinsic properties of points, and in one of them it is a law that all *F* s are *G*s, and in the other it is a law that all *G*s are *F*s, and there are further intrinsic properties of all the points which are *F* and *G* which underlie these laws without making a difference to any of the other facts. So we imagine that the property *being F in virtue of being G* is held by all these things in one world but not in the other, and this is a fundamental perfectly natural property. I don't think any of this is literally inconsistent, and I think filling out the details could give us a way for nomological reductionism to fail while the letter of Humean supervenience holds. But it would clearly violate the spirit of Humean supervenience and it isn't clear why we should believe in such "possibilities" anyway.

So in practice, I think that any philosopher who rejects nomological reductionism is probably going to want to reject Humean supervenience. And I think that Lewis saw some of the deepest challenges to Humean supervenience as coming from threats to nomological reductionism. In particular, Lewis thought that the biggest challenges to Humean supervenience came from the difficulties in providing a reductive account of chance, and the appeal of non-reductive series of causation.

The difficulties in providing a reductive account of chance are discussed at length in the introduction to Lewis (1986b), and in the only paper that has "Humean supervenience" in its title (Lewis 1994a). Here is a quick version of the problem. Chances are not fundamental, so they must supervene on the distribution of qualities. At least in the very early stages of the universe, there aren't enough facts about the distribution of qualities in the past and present to form a suitable subvenient base for the chances. So whether the chance of p is x or y will, at least some of the time, depend on how the future of the world turns out. Now let p the proposition that tells the full story about the future of the world. And assume that p is a proposition such that what its chance is depends on how that future goes. If it goes the way p says it will go, the chance of p is x ; if it goes some other way, the chance of p is y . Given a Humean theory of chance, Lewis says that this is going to be possible.

But now there's a problem. What Lewis calls the Principal Principle says that if we know the chance of p is y , and have no further information, then our credence in p should be y . But, in this case, if we knew the chance of p was y , we could be sure that p would not obtain. So our credence in p should be 0. Here we seem to have reached a contradiction, and it is a contradiction Lewis for a long time feared undermined the prospect of giving a reductive account of chance. The solution he eventually settled on in Lewis (1994a) was to slightly modify the Principal Principle, with the modification being designed to make very little difference in regular cases, but avoid this contradiction.

Lewis discusses the appeal of non-reductive theories of causation in several places, most notably for our purposes Lewis (2004a) and Lewis (2004c). Much of his attention is focused on the theory developed by Peter Menzies (1996). Menzies suggests that causation is the intrinsic relation that does the best job of satisfying folk platitudes about causation. A consequence of Menzies's view is that there is something that makes a difference to the intrinsic properties of pairs of causes and effects which doesn't supervene on either the intrinsic properties of the two ends of the causal chain, or on the spatiotemporal relations that hold between them. This something will either be causation or will be something on which causation depends. Either way there is a problem for Humean supervenience, since there will have to be a perfectly natural relation that is not spatiotemporal.

Lewis's response is to raise problems for the idea that causation could be an intrinsic relation. One class of worries concerns the very idea that causation could be a relation. Lewis says that absences can be causes and effects, but absences can't stand in any relations, so causation must not be a

relation. Another class of worries concerns the idea that causation could be intrinsic. Causation by double prevention, says Lewis, doesn't look like it could be intrinsic. But intuitively there could be causation by double prevention. Yet another class of worries concerns the idea that causation could be a natural relation, or that there could be any one thing that satisfies all the platitudes about causation. The vast array of different ways in which causes can bring about their effects in the actual world, he says, undermines this possibility.

(Note: In both cases Lewis defends Humean supervenience simply by defending nomological reductionism. So I think it is fair to say that there's a close connection between the two in Lewis's overall theory.)

8.5 Why Care about Humean Supervenience

As is well known, some surprising results in quantum mechanics suggest that entanglement relations are somehow fundamental (Maudlin 1994). This suggests that Humean supervenience is actually false. If that's right, why should we care about philosophical arguments for Humean supervenience? Lewis's response to this challenge is somewhat disconcerting.

> Really, what I uphold is not so much the truth of Humean Supervenience as the *tenability* of it. If physics itself were to teach me that it is false, I wouldn't grieve.
>
> That might happen: maybe the lesson of Bell's Theorem is exactly that . . . But I am not ready to take lessons in ontology from quantum physics as it now is. . . . If, after [quantum theory has been cleaned up], it still teaches non-locality, I shall submit willingly to the best of authority.
>
> What I want to fight are *philosophical* arguments against Humean Supervenience. When philosophers claim that one or another commonplace feature of the world cannot supervene on the arrangement of qualities, I make it my business to resist. Being a commonsensical fellow (except where unactualised possible worlds are concerned) I will seldom deny that the features in question exist. I grant their existence, and do my best to show how they can, after all, supervene on the arrangement of qualities. (Lewis 1986b, xi)

We can, I think, dismiss the point about quantum physics. The theory has been cleaned up in just the way Lewis wanted, and the claims about non-locality remain. Indeed, by the end of his life Lewis was willing to take lessons in ontology from quantum physics. See, for example, Lewis (2004b). So what is at issue here is whether or not there are philosophical arguments against Humean supervenience.

But at this point we might wonder why we should care. If a theory is false, what does it matter whether its falsehood is shown by philosophy or by physics? We might compare the dismissive attitude Lewis takes towards Plantinga's attempts to show that reconstructions of the problem of evil as an argument do not rely solely on things provable in first-order logic (Lewis 1993).

The answer I offered in Weatherson (2010) was that the philosophical defence of Humean supervenience was connected to the point of the last paragraph quoted above. Lewis wanted to save various features of our commonsensical picture of the world. And he wanted to do this without saying that philosophical reflection showed us that the picture of the world given to us by signs of somehow incomplete. He wanted to defend what I called "compatibilism," something that I contrasted with eliminativism and expansionism. The eliminativists want to say that science shows us that some commonsensical feature of reality doesn't really exist. (See, for example, Churchland (1981) for eliminativism about folk psychological states.) The expansionists want to say that since science (or at least physics) doesn't recognize certain features of reality, but they obviously exist, we need to posit

that science (or at least physics) is incomplete. There are many stripes of philosophical expansionists, from theists to dualists to believers in agent causation.

Lewis wasn't averse in principle to either eliminativism or expansionism. One could, depending on exactly how one interpreted folk theory and science, classify him as an eliminativist about gods, and an expansionist about unactualized possible worlds. But his first tendency was always to support compatibilism. Compatibilists face what Frank Jackson (1998) called the "location problem." They have to show where the commonsensical features are located in the scientific picture. That is, they have to show how to reduce (in at least some sense of "reduce") commonsensical concepts to scientific concepts. (Many compatibilists may bristle at the idea that they have to be reductionists; in recent decades the world has abounded with "non-reductive physicalists," who are precisely compatibilists in my sense, but who reject what they call "reductionism." But as Lewis (1994b) argued, these rejections often turn on reading too much into the notion of reduction. For that reason, Lewis would not have objected to being described as a reductionist about many everyday concepts.)

One way to perform such a reduction would be to wait until the best scientific theory is developed, and show where within it we find minds, meanings, morals and all the other exciting features of our ordinary world-view. But that could take a while, and philosophers could use something to do while waiting. In the meantime we could look for a recipe that should work no matter what physical theory the scientists settle on, or at least should work in a very wide range of cases. I think we can see Lewis's defense of Humean supervenience as providing such a recipe.

It is important to note here that Lewis's defense of Humean supervenience was largely *constructive*. He didn't try to give a proof that there couldn't be more to the world than the arrangement of local qualities. At least, he didn't rest a huge amount of weight on such arguments. The arguments we will look at below for a functional construal of the nomological are, perhaps, hints at arguments of this type. But, in general, Lewis defended Humean supervenience by explicitly showing where the ordinary concepts fitted in to a sparse physical picture of reality, under the assumption that physics tells us that the world consists of nothing but a spatiotemporal arrangement of intrinsic qualities.

Now physics tells us no such thing. But it shouldn't matter. If the recipe Lewis provides works in the case of the "Humean" world, it should also work in the world physics tells us we actually live in. The reduction of laws to facts about the distribution of fundamental qualities, and the reduction of chances and counterfactual dependencies to facts about laws, and the reduction of causation to facts about chances and counterfactual dependencies, and the reduction of mind to facts about causation and the distribution of qualities, and the reduction of value to facts about minds, and so on are all independent of whether physics tells us that we have to recognize relationships like entanglement as fundamental. In other words, if we can solve the location problem for the Humean world, we can solve it for the actual world. And solving the location problem is crucial to defending compatibilism. And whether it is possible to defend compatibilism is a central concern of metaphysics.

I quoted above a passage from 1986 in which Lewis links Humean supervenience to compatibilism. It's worth noting that he returns to the point in 1994.

> The point of defending Humean Supervenience is not to support reactionary physics, but rather to resist philosophical arguments that there are more things in heaven and earth than physics has dreamt of. Therefore if I defend the *philosophical* tenability of Humean Supervenience, that defence can doubtless be adapted to whatever better supervenience thesis may emerge from better physics. (Lewis 1994a, 474)

That is, the defense of Humean supervenience just is part of the argument against expansionism, and hence for compatibilism. That was the defense I offered in Weatherson (2010) for the interest of

Lewis's defense of Humean supervenience, even if it were to turn out that Humean supervenience was refuted by physics. I still think much of it is correct. In particular, I still think that Lewis wanted to defend compatibilism, and that the defense of Humean supervenience is key to the defense of compatibilism. Indeed, I think there is pretty strong textual evidence that it was a major part of Lewis's motivation for defending Humean supervenience. But this explanation of why the defense of Humean supervenience is significant can't explain why Lewis was so worried about the failures of Humean theories of chance. After all, if all we are trying to do is show that science and common sense are compatible, we could just take chances to be one of the fundamental features of reality given to us by science. There isn't any need, from the perspective of trying to reconcile science and common sense, to give a reductive account of chance. Yet Lewis clearly thought that giving a reductive account of chance was crucial to the defense of Humean supervenience. As he said,

> There is one big bad bug: chance. It is here, and here alone, that I fear defeat. But if I'm beaten here, then the entire campaign goes kaput. (Lewis 1986b, xiv)

I now think that attitude is very hard to explain if my early views about the significance of Humean supervenience are entirely correct. The natural conclusion is that there is something *more* that the defense of Humean supervenience is supposed to accomplish. One plausible interpretation is that what it is supposed to accomplish is a vindication of the idea that the key nomological concepts are, in a sense, descriptive. It's easiest to say what this sense is by contrasting it with the kind of view that Lewis rejected.

We're all familiar with the standard story about "water." Our ordinary usage of the term latches onto some stuff in the physical world. That stuff is H_2O. Some people think that's because our ordinary usage determines a property which H_2O satisfies, others because we demonstratively pick out H_2O in ordinary demonstrations of what it is we're talking about when we use the term "water." Either way, we get to be talking about H_2O when we use the word "water," even if we are so ignorant of chemistry that we can't tell hydrogen and oxygen apart. Moreover, our term continues to pick out "water" even in worlds that are completely free of hydrogen and oxygen, and even if such worlds have other stuff that plays a very similar functional role to the role water plays in the actual world.

Lewis was somewhat skeptical of this standard story about "water" (Lewis 2002). He thought that the ordinary term was ambiguous between our usage on which it picked out H_2O, and usage on which it picked out a role, a role that happens to be played by H_2O in the actual world but which could be played by other substances in other worlds. But if he thought the standard story about "water" was at best, part right, he thought applying a similar story to "law," "cause," and "chance" was wildly implausible.

If such a story were right, then we would expect to find worlds where there was some relation other than causation which played the causal role. Since the actual world is physical, any world in which non-physical things stand in the kind of relations that causes and effects typically stand in should do. So, for instance, if we have a world where the castings of spells are frequently followed by transformations from human to toad form, we should have a world where spells don't cause such transformations but rather the spellcasting and the transformation stand in a kind of fool's cause relationship. But we see no such thing. In such magical worlds, spells cause transformations.

So whatever causation is, it doesn't look to be the kind of thing whose essence can be discovered by physics. Physics couldn't tell us anything about the essence of the relationship between the spell and the transformation into a toad. But, we think, physics can tell us a lot about the fundamental properties and relations are instantiated in the actual world. So causation must not be one of them.

Lewis has a number of other arguments against anti-descriptivist views about individual nomological concepts. These arguments strike me as rather strong in the case of lawhood and causation, and less strong in the case of chance.

> If *being F* and *being F in virtue of a law* are both fundamental properties, then a plausible principle of modal recombination would suggest they could come apart. But they cannot; or at least they cannot in one direction. We want *being F in virtue of a law* to entail *being F*. That's easy if lawhood is defined in terms of fundamental properties of things; but it's hard to see how it could be if lawhood itself is fundamental (Lewis 1986b, xii).

A similar argument goes for causation. Assume that causation is a fundamental intrinsic relation that holds between things at different times. Consider, for instance, the causal relationship which holds between a throw of a rock (call it t) and the shattering of the window (call it s). As we noted above in the case of Daniels and O'Leary, several applications of modal combinatorialism suggest that there will be a world just like this one in which t is followed by s, but in which t does not cause s. But such a world seems to be impossible. As we also noted above, such a view runs into trouble with causation by double prevention, which does not look to be intrinsic.

The last two paragraphs have been extremely quick arguments, but in both cases it seems to me that they can be tightened up so as to provide good arguments for some kind of descriptivist stance towards laws and causation. Chance is another matter.

The first problem is that recombination arguments if anything point *away* from descriptivism about chance. Any such account will imply that chances can't, in general, point too far away from frequencies. But recombination arguments suggest that chances and frequencies can come arbitrarily far apart. Consider some particular event type e that has a one-half chance of occurring in circumstances c. Start with a world where c occurs frequently, and about half the time it is followed by e. Now use recombination to generate a world where all the $c \wedge \neg e$ events are deleted, so c is always followed by e. Unless we add a lot of bells and whistles to our theory of chance, it will no longer be the case that the chance of e given c is one-half. That is odd; we can't simply take the first circumstance where c occurred and at that moment there was a one-half chance of it being followed by e, and patch it into an arbitrary world. Bigelow, Collins and Pargeter (1993) turn this idea into a more careful argument against descriptivism about chance. They say that chances should satisfy the following principle. (In this principle, *Ch* is the chance function, and various subscripts relativize it to times and worlds.)

> Suppose $x > 0$ and $\mathrm{C}\,\mathrm{h}_{t\,w}(A) = x$. Then A is true in at least one of those worlds w' that matches w up to time t and for which $\mathrm{C}\,\mathrm{h}_{t}(A) = x$. (Bigelow *et al.* 1993, 459)

That is, if the chance of A at t is x, and $x > 0$, then A could occur without changing the history prior to t, and without changing the chance of A at t. This seems like a plausible principle of chance, but it entails the not-so-Humean view that chances at t supervene on history to t, not on the full state of the world.

Now as it turns out Lewis *doesn't* rest on recombination arguments against rival views of chance, and in my view he is wise to do so. Instead he rests on epistemological arguments. He takes the following two things to be data points.

1. Something like the Principal Principle is true. The original Principal Principle said that if you knew the chance of p at t was x, and didn't have any "inadmissible" information (roughly, information about how the world developed after t), then your credence in p should be x. Lewis

tinkered with this slightly, as we noted above, but he took it to be a requirement of a theory of chance that the Principal Principle turn out at least roughly right.

2. The correct theory of chance will *explain* the Principal Principle.

Lewis frequently wielded this second requirement against rival theories of chance. Here's one example.

> I can see, dimly, how it might be rational to conform my credences about outcomes to my credences about history, symmetries and frequencies. I haven't the faintest notion how it might be rational to conform my credences about outcomes to my credences about some mysterious unHumean magnitude. Don't try to take away the mystery by saying that this unHumean magnitude is none other than chance! (Lewis 1986b, xv)

But this also seems like a weak argument. For one thing, chances are actually correlated very well with frequencies, and this correlation does not look at all accidental. It seems very plausible to me that we should line up our credences with things that are actually correlated well with frequencies. But, you might protest, shouldn't we have an explanation of why the Principal Principle is an *a priori* principle of rationality? I think that before we ask for such an explanation, we should check how confident we are that the Principal Principle, or anything else, is part of an *a priori* theory of rationality. I'm not so confident that we'll be able to do this (Weatherson 2005, 2007).

There are other replies too that we might make. It seems plausible that we should minimize the expected inaccuracy of our credences (Joyce 1998). This is true when we consider not just the *subjective* expected inaccuracy of our credences, but the *objective* expected inaccuracy of our credences. That is, when we calculate the expected inaccuracy of someone's credences, using chances as the probabilities for generating the expectations, it is good if this expected inaccuracy is as low as possible. But, assuming that we are using a proper scoring rule for measuring the accuracy of credences, this means that we must have credences match chances.

More generally, I'm very skeptical of theories that insist our metaphysics be designed to have complicated epistemological theses fall out as immediate consequences. Rationality requires that we be inductivists. Why is that? Here's a bad way to go about answering it: find a theory of persistence that makes induction obviously rational, and then require our metaphysics to conform to that theory. I don't think you'll get a very good theory of persistence that way, and, relatedly, you won't get a very Lewisian theory of persistence that way. The demand that the theory of chance play a central role in an explanation of the Principal Principle strikes me as equally mistaken.

If what I've been saying so far is correct, then chance interacts with the motivation for Humean supervenience in very different ways to how laws and causation interact. Neither of the two kinds of motivations for defending Humean supervenience against philosophical attacks provides us with good reason to leave chances out of the subvenient base on which we say all contingent facts supervene. This is not to yet offer anything like a positive argument for chances to be part of the fundamental furniture of reality. Rather, what I've argued here is that a metaphysics that takes chances as primitives would not be as far removed from a recognizably Lewisian metaphysics as a metaphysics that takes laws or causes as primitive, let alone one that takes mind, meanings or morals as primitive.

8.6 Points, Vectors, and Lewis

The other main point from the discussion of the previous section is that the fact that quantum mechanics raises problems for Humean supervenience does not undercut the philosophical

significance of Lewis's defense of Humean supervenience. But is Humean supervenience even compatible with classical physics? Perhaps not.

> Even classical electromagnetism raises a question for Humean Supervenience as I stated it. Denis Robinson (1989) has asked: is a vector field an arrangement of local qualities? I said qualities were intrinsic; that means they can never differ between duplicates; and I would have said offhand that two things can be duplicates even if they point in different directions. Maybe this last opinion should be reconsidered, so that vector-valued magnitudes may count as intrinsic properties. What else could they be? Any attempt to reconstruct with them as relational properties seems seriously artificial. (Lewis 1994a, 474)

The opinion that Lewis proposes to discard here seems more than an offhand judgment. It seems to follow from the very way that we introduce the notion of duplication. Here is Lewis's own attempt to introduce the notion.

> We are familiar with cases of approximate duplication, e.g. when we use copying machines. And we understand that if these machines were more perfect than they are, the copies they made would be perfect duplicates of the original. Copy and original would be alike in size and shape and chemical composition of the ink marks and the paper, alike in temperature and magnetic alignment and electrostatic charge, alike even in the exact arrangement of their electrons and quarks. Such duplicates would be exactly alike we say. They would match perfectly, they would be qualitatively identical, they would be indiscernible. (Lewis 1983, 355)

If Lewis is right that vector-valued magnitudes may count as intrinsic properties, then there is yet another condition that the perfect copying machine must satisfy. The original and the duplicate must be parallel. This isn't the case in most actual copying machines. Usually, the original is laid flat, while the duplicate is a small angle to make it easier to collect. This is a feature, not a bug. It is not a way in which the machine falls short of perfect copying. But if vector-valued magnitudes are intrinsic qualities, and duplicates share their intrinsic qualities, it would be. So Lewis is wrong to think that these vector-valued magnitudes may be intrinsic.

Moreover, the little argument that Lewis gives seems to rest on a category mistake. What matters here is the division of properties into intrinsic and extrinsic. But properties are not the kind of things that can be relational or non-relational. As Humberstone (1996) shows, it is concepts not properties that are relational and non-relational. For instance the concept *being the same shape as David Lewis actually was at noon on January 1, 1970*, is a relational concept that picks out an intrinsic shape property. Whether vector-valued magnitudes are intrinsic or extrinsic properties is independent of the question of whether it is best to pick them out by means of relational or non-relational concepts.

There is a further issue about the compatibility of Humean supervenience with classical physics. This is a point that has been made well by Jeremy Butterfield (2006), and we can see the problem by looking at the different ways in which Lewis introduces Humean supervenience.

> Humean Supervenience says that in a world like ours, the fundamental properties are local qualities: perfectly natural intrinsic properties of points, or of point-sized occupants of points. (Lewis 1994a, 474)

Lewis goes back and forth between local properties and intrinsic properties of points here. These aren't the same thing. As Butterfield notes, "local" is used in a few different ways throughout physics. One simple usage identifies local properties of a point with properties that supervene on intrinsic features of arbitrarily small regions around the point. To take an important example, the slope of a curve at a point may be a local property of the curve at that point without being intrinsic property of the point.

This raises a question: can we do classical physics with only intrinsic properties of points, and not even these further local properties? Butterfield argues, persuasively, that the answer is no. He notes, however, that there are some very mild weakenings of Humean supervenience that avoid this difficulty. Here is a very simple one.

Call local supervenience the following thesis. For any length ε greater than 0, there is a length d less than ε with the following feature. All the facts about the world supervene on intrinsic features of objects and regions with diameter at most d, plus facts about the spatiotemporal arrangement of these objects and regions. This will mean that we can include all local qualities in the subvenient base, without assuming that these are intrinsic qualities of points. If the theory of intrinsicness in Weatherson (2006) is correct, we'll also be able to include vector-valued magnitudes in the subvenient base without assuming that these are intrinsic properties of points. (On my view, they will end up being intrinsic properties of asymmetrically shaped regions.) We still won't be able to accommodate entanglement relationships, but we will be able to capture classical physics. And, for the reasons discussed in the previous section, it would still be worthwhile to ask whether there are philosophical objections to local supervenience. A negative answer would greatly assist the arguments for compatibilism, and for nomological descriptivism.

Butterfield offers theses like local supervenience to Lewis as friendly suggestions. But he thinks Lewis's focus on points and their properties would have led him to reject it. I don't want to get into the business of making counterfactual speculations about what Lewis would or would not have accepted. But I think he should have been happy to weaken Humean supervenience to something like local supervenience. If the point of defending Humean supervenience is not to defend its truth, but rather to assist in larger arguments for compatibilism, and for nomological descriptivism, then the big question to ask is whether a defense of local supervenience (against distinctively philosophical objections) would have served those causes just as well. And I think it's pretty clear that it would have. Showing that we have no philosophical reason to posit fundamental non-local features of reality would be enough to let us "resist philosophical arguments that there are more things in heaven and earth than physics has dreamt of " (Lewis 1994a, 474). Lewis's work in defending Humean supervenience has been invaluable to those of us who want to join this resistance. It wouldn't have been undermined if he'd allowed some local properties into the mix.

References

Bigelow, J. (1988). *The reality of numbers: a physicalist's philosophy of mathematics*. Oxford: Oxford University Press.

Bigelow, J., Collins, J. & Pargetter, R. (1993). The big bad bug: what are the Humean's chances? *British Journal for the Philosophy of Science*, 44, 443–62.

Butterfield, J. (2006). Against pointillisme about mechanics. *British Journal for the Philosophy of Science*, 57, 709–53.

Churchland, P. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy*, 78, 67–90.

Goodman, N. (1955). *Fact, fiction and forecast*. Cambridge, MA: Harvard University Press.

Hall, N. (2010). David Lewis's metaphysics. In E.N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2010 edition), http://plato.stanford.edu/entries/lewis-metaphysics/, accessed June 19, 2014.

Haslanger, S. (1994). Humean supervenience and enduring things. *Australasian Journal of Philosophy*, 72, 339–59.

Humberstone, I.L. (1996). Intrinsic/Extrinsic. *Synthese*, 108, 205–67.

Jackson, F. (1998). *From metaphysics to ethics: a defence of conceptual analysis*. Oxford: Clarendon Press.

Joyce, J.M. (1998). A non-pragmatic vindication of probabilism. *Philosophy of Science*, 65, 575–603.

Lewis, D. (1973). Causation. *Journal of Philosophy*, 70, 556–7. Reprinted in Lewis (1986b), pp. 159–72.
Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, 61, 343–77.
Lewis, D. (1986a). *On the plurality of worlds*. Oxford: Blackwell.
Lewis, D. (1986b). *Philosophical papers*, Vol. II. Oxford: Oxford University Press.
Lewis, D. (1993). Evil for freedom's sake? *Philosophical Papers*, 22, 149–72.
Lewis, D. (1994a). Humean supervenience debugged. *Mind*, 103, 473–90. Reprinted in Lewis (1999), pp. 224–47.
Lewis, D. (1994b). Reduction of mind. In S. Guttenplan (Ed.), *A companion to the philosophy of mind*, pp. 412–31. Oxford: Blackwell.
Lewis, D. (1999). *Papers in metaphysics and epistemology*. Cambridge: Cambridge University Press.
Lewis, D. (2002). Tharp's third theorem. *Analysis*, 62, 95–7.
Lewis, D. (2004a). Causation as influence. In J. Collins, N. Hall & L.A. Paul (Eds.), *Causation and counterfactuals*, pp.75–106. Cambridge, MA: MIT Press.
Lewis, D. (2004b). How many lives has Schrödinger's cat? *Australasian Journal of Philosophy*, 82, 3–22.
Lewis, D. (2004c). Void and object. In J. Collins, N. Hall & L.A. Paul (Eds.), *Causation and counterfactuals*, pp.277–90. Cambridge, MA: MIT Press.
Maudlin, T. (1994). *Quantum non-locality and relativity: metaphysical intimations of modern physics*. Oxford: Blackwell.
Menzies, P. (1996). Probabilistic causation and the pre-emption problem. *Mind*, 105, 85–117.
Nolan, D. (1996). Recombination unbound. *Philosophical Studies*, 84, 239–62.
Pryor, J. (2000). The sceptic and the dogmatist. *Noûs*, 34, 517–49.
Robinson, D. (1989). Matter, motion and Humean supervenience. *Australasian Journal of Philosophy*, 67, 394–409.
Schwarz, W. (2009). *David Lewis: metaphysik und analyse*. Paderborn: Mentis-Verlag.
Sider, T. (1993). Naturalness, intrinsicality and duplication. PhD thesis, University of Massachusetts–Amherst.
Sider, T. (2001). Criteria of personal identity and the limits of conceptual analysis. *Philosophical Perspectives*, 15, 189–209.
Strawson, G. (2000). David Hume: objects and power. In R. Read & K.A. Richman (Eds.), *The new Hume debate*, pp. 31–51. London: Routledge.
Weatherson, B. (2003). What good are counterexamples? *Philosophical Studies*, 115, 1–31.
Weatherson, B. (2005). Scepticism, rationalism and externalism. *Oxford Studies in Epistemology*, 1, 311–31.
Weatherson, B. (2006). The asymmetric magnets problem. *Philosophical Perspectives*, 20, 479–92.
Weatherson, B. (2007). The Bayesian and the dogmatist. *Proceedings of the Aristotelian Society*, 107, 169–85.
Weatherson, B. (2010). David Lewis. In E.N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2010 edition), http://plato.stanford.edu/archives/sum2010/entries/david-lewis/, accessed May 28, 2014.

9

No Work for a Theory of Universals

M. EDDON AND C.J.G. MEACHAM

9.1 Introduction

In "New Work for a Theory of Universals" (1983), David Lewis proposes a version of the Best System account of lawhood. On Lewis's account, a law is a sentence entailed by the best system, where the best system is the set of true sentences that best balances simplicity and informativeness, and whose predicates refer to only the perfectly natural properties and relations.[1]

Many philosophers find the Best System account of lawhood appealing, at least in broad outline.[2] But Lewis's version of the Best System account requires a primitive distinction between properties that are *perfectly natural* and those that are not. And those who find a Best System account attractive because it does not posit mysterious or inscrutable primitive causal powers, governing relations, and the like, tend to be unhappy with this appeal to perfectly natural properties.[3] For, one might say, the distinction Lewis relies on is no less mysterious or inscrutable.

Is there a way to develop a Best System account of lawhood that preserves the features that these philosophers find appealing, and does not invoke a primitive distinction among properties? Although a number of people have discussed this question, the literature on this topic is somewhat disjointed. Different parties focus on different issues, and set up the dialectic in different ways.

This paper has three goals. The first is to provide a framework for the debate. We lay out the space of available positions on this question, pinpoint the crucial junctures, and show how participants in the debate can be seen as taking different stances at these junctures.

The second goal is to evaluate the prospects of these different positions, and to identify the most promising candidates. After surveying the options, we argue that there are reasons to prefer a version of the Best System account of lawhood suggested by Loewer (2007) and Hoefer (2007) over others.

The third goal is to broaden the horizons of the original project. Perfectly natural properties appear not only in Lewis's analysis of lawhood, but also in his accounts of objective resemblance, duplication, intrinsicality, causation, counterfactuals, certain supervenience claims, and reference. Is there

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

a way to modify these accounts, so that they too can do without perfect naturalness? We suggest one way to do this using the version of the Best System account discussed above.

The paper proceeds as follows. In section 9.2, we present the basic form of the Best System account, and sketch some ways this account may be developed. Using this framework, we lay out the landscape of proposals that people have offered. In section 9.3, we discuss some of the worries facing the different proposals, and select the options that strike us as the most promising. In section 9.4, we consider the question of whether one can modify Lewis's other accounts as well, so that they too do not appeal to perfectly natural properties. We develop one proposal, and consider some objections. Finally, in section 9.5, we briefly consider a different route one might take, drawing on work by Williams (chapter 23, this volume).

9.2 Best System Accounts of Laws

Let's begin by sketching a basic template for a Best System account of lawhood. To keep things simple, we will generally restrict our attention to non-probabilistic laws.

Consider the sets of true sentences. Some of these sets are more informative than others – they tell us more about what the world is like. Some of these sets are simpler than others – they contain fewer or shorter sentences. These two features tend to compete with one another. Adding more sentences to a set may increase its informativeness, but decreases simplicity; removing sentences may decrease informativeness, but increases simplicity. Now consider the set of true sentences that does best, on balance, at satisfying the desiderata of simplicity and informativeness. Call this set the *best system*. Anything entailed by the best system is a law of nature.[4]

This template is silent on two key questions. The first question has to do with the language in which sentences of candidate systems are formulated.[5] Consider Lewis's example of a language containing the predicate F, where F applies to all and only the things at the actual world.[6] The set containing only the sentence $\forall x \mathrm{F}x$ is maximally simple (it contains only one short sentence) and maximally informative (it entails every truth about the actual world). If it is permissible to formulate systems in a language containing F, then we get the absurd result that $\{\forall x \mathrm{F}x\}$ is the best system, and every truth is a law. In order to rule out a system like this, we need to place some constraints on the languages employed by candidate systems. So what language should be used to formulate candidate systems? Call this *the Language Question*.

The second question has to do with our evaluation of how well candidate systems satisfy the relevant desiderata. There are a number of ways one might assess the simplicity or informativeness of a set of sentences, as well as how these two desiderata balance off against one another. One metric of simplicity might take some logical operators to add more to a sentence's complexity than others – maybe material implication adds more complexity than negation. One metric of informativeness might take certain sentences to be more informative than others, even though neither sentence entails the other – maybe the sentence describing the future history of the universe is more informative than the sentence describing what I ate for breakfast. And different metrics of balance might weigh simplicity against informativeness in different ways. Thus, which system is best, and *a fortiori* what the laws are, may vary depending on the metrics of simplicity, informativeness, balance, and whatever else we deem relevant. So which metrics should we use to evaluate candidate systems? Call this *the Metrics Question*.

Different answers to these questions yield different Best System accounts. In the following sections, we briefly describe the answers to these questions that have been offered in the literature. We then sketch the landscape of the literature, laying out the space of positions taken by various authors.

9.2.1 Answers to the Language Question

The Language Question: What language should we use to formulate candidate systems?

Language-Objectivity: An Objective, "Mind-Independent" Language
This is Lewis's preferred answer to the Language Question; on his view, the language of candidate systems is the one containing only those predicates that refer to the perfectly natural properties. This language is objective or "mind-independent" because which properties are perfectly natural is an objective matter, having nothing to do with us.

Language-Indexicality: The Language of the Speaker
According to Language-Indexicality, when one asserts that *P* is a law, one's claim is true iff, when the candidate systems are formulated in one's language, the best system entails *P*. (Alternatively, one might hold that the appropriate language of candidate systems is the language of the society to which the person belongs, or perhaps the language of that society's scientists.)

Language-Rigidity: Our Language
According to this answer, we take the language of candidate systems to be *our* language. (Or, alternatively, the language of our society, or our society's scientists, or some idealized version of our society's scientists.) Furthermore, when considering candidate systems of other possible worlds, we formulate these systems in our actual language – not the language we would have had at the world under consideration. In other words, we "rigidify" our language.

Language-Relativity: Any Language – and Best Systemhood Is Language-Relative
A fourth option is to allow any language to be used to formulate candidate systems, and make "best-systemhood" – and hence lawhood – relative to a language. So, for any language *L*, candidate systems-in-*L* vie for best system-in-*L*. Consider two languages, L_1 and L_2. Suppose that *P* is entailed by the best system-in-L_1 at world *w*, and *Q* is entailed by the best system-in-L_2 at *w*. Then *P* is a law-in-L_1 at *w*, and *Q* is a law-in-L_2 at *w*. But neither *P* nor *Q* is a law *simpliciter*, since there are no laws *simpliciter*.

Language-Salience: All Languages – and the Best System Must Be Salient
A fifth option is to push the burden of choosing a language elsewhere. Here is one way to do that. Consider all candidate systems, formulated in all possible languages. In choosing among candidate systems, we consider not only simplicity and informativeness, but also whether a system is unified, useful, and explanatory – for short, *salient*. On the Language-Salience answer, the language of the best system is the language of the system that best balances simplicity, informativeness, and salience. Adding this desideratum allows us to rule out systems such as $\{\forall x \mathrm{F}x\}$; for while $\{\forall x \mathrm{F}x\}$ scores highly on simplicity and informativeness, it scores very low on salience.[7,8]

9.2.2 Answers to the Metrics Question

The Metrics Question: What metrics of simplicity, informativeness, balance, and whatever else we deem relevant, should we use when evaluating candidate systems?

Metrics-Objectivity: The Objective, "Mind-Independent" Metrics
The Metrics-Objectivity answer says that there is an objectively correct way to measure the extent to which a system satisfies the specified criteria (e.g., simplicity, informativeness, balance, etc.).

Metrics-Indexicality: The Metrics of the Speaker
According to Metrics-Indexicality, when one asserts that *P* is a law, one's claim is true iff, when candidate systems are assessed using the metrics of the speaker, the system judged best entails *P*. (Alternatively, one might hold that candidate systems should be evaluated using the metrics of the society to which the person belongs, or perhaps the metrics of that society's scientists.)

Metrics-Rigidity: Our Metrics
Another option is to take the metrics used to evaluate the extent to which candidate systems satisfy the specified criteria to be *our* metrics. (Or, alternatively, the metrics of our society, or our society's scientists, or some idealized version of our society's scientists.) As with the Language-Rigidity answer, "our metrics" should be understood rigidly – when considering candidate systems of other possible worlds, we assess these systems using our actual metrics, not those we would have had at the world under consideration.

Metrics-Relativity: Any Metrics – and Best Systemhood Is Metrics-Relative
On the Metrics-Relativity answer, we evaluate candidate systems using any metrics, and make "best-systemhood" relative to the metrics employed. So, for any set of metrics *M*, there is a best system-evaluated-by-*M*. Consider two sets of metrics, M_1 and M_2. Suppose that *P* is entailed by the best system-evaluated-by-M_1 at *w*, and *Q* is entailed by the best system-evaluated-by-M_2 at *w*. *P* is a law-evaluated-by-M_1, and *Q* is a law-evaluated-by-M_2. But neither is a law *simpliciter*, since there are no laws *simpliciter*.

9.2.3 The Landscape

Table 9.1 summarizes the different positions that authors have taken in recent literature. Because some authors defend positions that are compatible with multiple answers, their names appear more than once.

Table 9.1 The different positions that authors have taken in recent literature

		Metrics Question			
		Objectivity	***Indexicality***	***Rigidity***	***Relativity***
Language Question	***Objectivity***	Lewis (1994)	Lewis (1994)	Lewis (1980) Lewis (1994)	Lewis (1994)
	Indexicality		Roberts (1999) Halpin (2003)		
	Rigidity	Loewer (2007)		Hoefer (2007) Loewer (2007)	
	Relativity	Taylor (1993)			Taylor (1993) Halpin (2003) Callender & Cohen (2009)
	Salience	Loewer (2007)		Hoefer (2007) Loewer (2007)	

Barry Taylor (1993) and Craig Callender and Jonathan Cohen (2009) both defend the Language-Relativity answer to the Language Question, though they differ on points of detail. Their answers to the Metrics Question are less straightforward. Callender and Cohen do not explicitly endorse any answer to the Metrics Question, though Metrics-Relativity seems to best fit the spirit of their proposal.[9] And Taylor seems open to both Metrics-Relativity and Metrics-Objectivity.[10] In general, however, all of them take the laws to be relative, and all of them take there to be no deep sense in which the laws relative to one relatum are privileged over the laws relative to some other relatum.

John F. Halpin (2003) offers what he calls a "contextual" or "perspectival" account of lawhood. On this account, lawhood is relativized to perspectives, and the appropriate perspective is selected by context.[11] He can be read as endorsing either the Relativity answers to both questions or the Indexicality answers to both questions.[12]

John Roberts (1999) suggests that the best way to develop a Best System analysis of lawhood invokes Metrics-Indexicality.[13] Although he does not directly address the Language Question, it seems to be in the spirit of his account to endorse Language-Indexicality as well. Thus Roberts's discussion suggests the Indexicality answers to both questions.

Barry Loewer (2007) suggests two answers to the Language Question. One is Language-Salience: we consider candidate systems in all possible languages, and add a desideratum which takes into account other scientific virtues besides simplicity and informativeness – such as usefulness and explanatory value.[14] The other is Language-Rigidity: the language of candidate systems is some suitably idealized version of the actual language of our scientific community.[15] As for the Metrics Question, Loewer believes that our standards of simplicity, informativeness, and salience are deeply rooted in our scientific tradition – what *we* find explanatory, useful, and so on. But he also believes that it's plausible that these notions may be at least partially objective.[16] This suggests some combination of Metrics-Rigidity and Metrics-Objectivity: in cases where the objective metrics do not select a unique candidate system as best, we appeal to those of the actual scientific community.

Although Carl Hoefer (2007) focuses on a Lewisian account of chance, rather than a Lewisian account of laws, his position fits neatly into this framework. Hoefer defends a position similar to Loewer's, adopting Metrics-Rigidity and some combination of Language-Salience and Language-Rigidity. But while Hoefer's treatment of salience is similar to Loewer's in spirit, it differs in letter. Instead of treating salience as an additional desideratum, Hoefer incorporates salience directly into the simplicity and informativeness requirements. Thus he suggests we understand the (salient)-simplicity desideratum as selecting for something like user-friendliness, taking comprehensibility and ease of use for agents like us into consideration. And he suggests that we understand the (salient)-informativeness desideratum as selecting for something like helpfulness, taking utility and relevance to agents like us into consideration.[17]

Finally, David Lewis endorses Language-Objectivity – appealing to perfectly natural properties to determine the language of candidate systems (Lewis 1983, 1994). Lewis's position on the Metrics Question is less clear. In earlier work, Lewis (1980) adopts Metrics-Rigidity: we employ *our* metrics when assessing candidate systems, and we hold these fixed when assessing systems at other possible worlds.[18] But he later revises his position.[19] His remarks in Lewis (1994) suggest two different responses to the Metrics Question. One option is Metrics-Objectivity. Lewis writes:

> I suppose our standards of simplicity and strength and balance are only partly a matter of psychology . . . maybe some of the exchange rates between aspects of simplicity, etc. are a psychological matter, but not just anything goes. If nature is kind, the best system will be *robustly* best – so far ahead of its rivals that it will come out first under any standards of simplicity and strength and balance. ([1994] 1999, 232–3)

And if nature is unkind, and no system is clearly best, then "the theorems of the barely-best system would not very well deserve the name of laws" (233). This amounts to the Objectivity answer to the Metrics Question, since only the objective metrics play a role in determining the laws. If nature is kind, then the objective metrics select the unique best system; if nature is unkind, and the objective metrics do not select a unique best system, then there are no laws.

The other option suggested by Lewis is a combination of Metrics-Objectivity with some other answer to the Metrics Question. Lewis writes: "I can admit that *if* nature were unkind, and *if* disagreeing rival systems were running neck-and-neck, then lawhood might be a psychological matter, and that would be very peculiar" ([1994] 1999, 233). So, at a world where the objective metrics do not select one system as clearly best, then the metrics we use to evaluate candidate systems would be partly determined by objective evaluative facts, and partly determined by facts about our psychology. This option amounts to some combination of Metrics-Objectivity with one of the other answers.[20]

9.3 Assessing the Options

Proponents of Best System accounts of laws have offered a number of different answers to the Language and Metrics Questions. But how do these different answers rate against one another? In this section, we discuss the challenges facing each of these answers, and offer our assessment of the options.

9.3.1 Language-Objectivity and Metrics-Objectivity

Let's begin with the two Objectivity answers. According to Language-Objectivity, there is an objectively correct, "mind-independent" language in which candidate systems are formulated. According to Metrics-Objectivity, there is an objectively correct, "mind-independent" set of metrics that determines how candidate systems score with respect to their balance of simplicity and informativeness (and whatever else).

It seems likely that those who are skeptical of Lewis's distinction between natural and non-natural properties will be skeptical of both the Objectivity answers as well. For appealing to an objectively correct language or set of metrics is (arguably) no less mysterious than appealing to a primitive distinction between natural and non-natural properties. Since this paper is addressed to those disinclined to adopt natural properties, we will put both Language-Objectivity and Metrics-Objectivity aside.

9.3.2 Language-Indexicality and Metrics-Indexicality

The Indexicality answers take the term "law" to be an indexical, like "here" and "now." But whereas the referents of "here" and "now" are determined by the spatial and temporal location of the speaker, the referent of "law" is determined by the language and metrics of the speaker.[21,22] For simplicity, our discussion focuses on Metrics-Indexicality, though what we say applies to both Indexicality answers.[23]

There are several worries facing the Indexicality answers. One worry is that, if the term "law" is an indexical, then there is not enough room for meaningful disagreement about lawhood.[24] Suppose Aristotle and Newton are having a conversation about what the laws of nature are. And suppose that Aristotle and Newton speak the same language, but have different metrics of simplicity, informativeness, and balance: the best system indexed to Aristotle's metrics yields Aristotelian Mechanics,

whereas the best system indexed to Newton's metrics yields Newtonian Mechanics. It seems that Aristotle and Newton should be able to meaningfully disagree about what the laws are. But suppose Aristotle says, "The laws are Aristotelian Mechanics," and Newton replies by saying, "The laws are Newtonian Mechanics." Given the Indexicality answers, this is no more of a disagreement than Aristotle saying, "I am at the North Pole," and Newton saying, "I am in Massachusetts." For when Aristotle says, "The laws are Aristotelian Mechanics," the term "law" refers to those given by Aristotelian Mechanics, whereas when Newton says, "The laws are Newtonian Mechanics," the term "law" refers to those given by Newtonian Mechanics. Aristotle and Newton are not disagreeing; they are just talking past each other.

Another worry for taking "law" to be an indexical term concerns the truth-values of counterfactuals involving embedded claims. Consider again Aristotle and Newton. If Aristotle says:

(1) If I were at the North Pole and I were to say, "It's cold here," then my assertion would be true,

then what he says is true. It does not matter whether Aristotle is actually in the Sahara Desert. The claim, "It's cold here," is evaluated with respect to the world at which he's at the North Pole; and at that world, it's cold at the North Pole. Likewise, treating "law" as an indexical like "here," if Aristotle says:

(2) If I had Newton's metrics and I were to say, "The laws are Newtonian Mechanics," then my assertion would be true,

then what he says is true. For the claim, "The laws are Newtonian Mechanics," is evaluated at the world where Aristotle has Newton's metrics; and at that world, his utterance is true. Intuitively, however, counterfactuals like (2) are false. But the Metrics-Indexicality answer entails that (2) is true.

One might try to avoid this result by maintaining that embedded claims in counterfactuals are evaluated using the speaker's actual metrics. But if one makes this move, then one cannot truly assert certain consequences of the Indexicality view. For instance, if Newton says "The laws are Newtonian Mechanics," then his assertion is true. But if Aristotle says:

(3) If Newton says "The laws are Newtonian Mechanics," then his assertion is true,

then *his* assertion is false. For on the present suggestion, we evaluate the embedded claim, "The laws are Newtonian Mechanics," using Aristotle's actual metrics. And Aristotle's metrics entail that the laws are not Newtonian Mechanics. So even though it is true that if Newton says "The laws are Newtonian Mechanics," then his assertion is true, this move entails that many of us who don't share Newton's metrics are unable to *say* that it's true. That seems wrong – surely those who don't share Newton's metrics can still truly assert (3). (Indeed, we've done so ourselves in the third sentence of this paragraph!)

The fact that the Indexicality answers deliver counterintuitive results for certain counterfactual claims reveals one way in which these answers appear to make the laws dependent on us in some important sense. But there is an even more straightforward way in which the Indexicality answers make the laws of nature us-dependent: if "law" is an indexical term, then the truth-values of sentences concerning the laws vary depending on who the speaker is. And that is a strange result. For whatever we say about what the laws are, it seems that they aren't the sorts of things that should vary from speaker to speaker.

9.3.3 Language-Rigidity and Metrics-Rigidity

According to the Language-Rigidity and Metrics-Rigidity answers, we use our language and metrics, understood rigidly, to determine the laws.[25,26] These answers can be understood in either a narrow way – as picking out the language and metrics of a particular individual – or a broad way – as picking out the language and metrics of an individual's scientific community. Since proponents of Rigidity answers generally adopt a broad understanding, that is what we assume in what follows.

Begin with the Language-Rigidity answer, according to which candidate systems are formulated in the language of our scientific community. This answer has some unpalatable consequences. The language of the scientific community today differs substantially from that of the scientific community 1000 years ago. And it seems likely that in 1000 years from now, the scientific community will be speaking of properties presently unknown to us, and using unfamiliar terminology from fields of mathematics yet to be developed. And so it seems reasonable to expect that the language of future scientists will have little in common with our current language, and that the laws posited by future scientists will be formulated in a language quite different from the one we currently use. If that's the case, and candidate systems are to be formulated in the language of our current scientific community, then future scientists are likely to get the laws wrong. But we generally think that future scientists are in a *better* position to discern the laws of nature, not a worse position.[27]

To avoid this result, an advocate of Language-Rigidity might say that the language of a candidate system is not our language, but rather an *ideal* version of our language: the language that an idealized version of our scientific community, in possession of all the non-nomic facts, would use to formulate their final scientific theory.[28] But this appeal to the ideal language and, by extension, the idealized scientific community, raises other worries.

One worry arises from the counterfactual nature of this approach. In order for L to be the ideal language, something like the following counterfactual must be true:

> **(4)** If our scientific community were idealized, and were to know all the non-nomic facts, then they would use L to formulate their final scientific theory.

But the truth value of such counterfactuals is generally taken to depend on what the laws are.[29] So this appeal to the ideal language threatens to make the account circular.

Another worry arises from the characterization of the ideal language as the one used by the idealized scientific community. The language an idealized scientific community will use to formulate their final theory will be one that makes their final theory attractive – simple, unified, perspicuous, and so on. If it's legitimate for our account of laws to appeal to the language this idealized scientific community would use to construct their final theory, then it should be legitimate for our account to appeal to the final theory this idealized scientific community would construct. But then we might as well appeal to the final theory directly: let the laws be whatever the final theory says they are.

But surely this is not a legitimate move. Appealing to an idealized scientific community just shifts the burden from describing how to get the laws from the non-nomic facts to describing how an idealized scientific community would construct their final theory from the non-nomic facts. The task in both cases is essentially the same – characterizing a function that takes the non-nomic facts as input, and spits out an attractive final theory as output. And re-describing the task in terms of idealized scientific communities doesn't make it any easier.

Now let's turn to Metrics-Rigidity. Like Language-Rigidity, this answer has an "actual" understanding (rigidify our current, actual metrics) and an "ideal" version (rigidify those of an ideal scientific community). The idealized version of Metrics-Rigidity runs into the same troubles as the

idealized version of Language-Rigidity. The non-idealized version of Metrics-Rigidity, on the other hand, seems to be on firmer ground than the non-idealized version of the Language-Rigidity answer. For while it's plausible that the language of the future scientific community will bear little resemblance to our current language, it's less plausible that the metrics of the future scientific community will bear little resemblance to our metrics of simplicity, informativeness, balance, etc. And thus the worry that arose for the non-idealized version of Language-Rigidity – that future scientists are likely to get the laws wrong because their language will differ from ours – does not seem as pressing for the non-idealized version of Metrics-Rigidity.

A more general worry, applicable to both the Language-Rigidity and Metrics-Rigidity answers, is that they threaten to make the laws of nature depend on us in some sense. For in appealing to our language and metrics, the Rigidity answers imply that the laws depend on the language and metrics of our scientific community. But intuitively, the laws of nature should not depend on what the language and metrics of any particular scientific community are like.[30]

This worry is hard to make precise, however.[31] One natural way to assess whether the laws depend on us – our language and metrics – is to apply a counterfactual test: the laws depend on us iff had our language and metrics been appropriately different, the laws would have been different as well. But on this construal of "us-dependence," the Rigidity answers do not imply that the laws depend on what we are like.[32] So to press this worry, one needs some other way of understanding the manner in which the Rigidity answers are "us-dependent."[33] Without that, it is unclear that this is a mark against the Rigidity answers.

In any case, the Rigidity answers seem better off than the Indexicality answers with respect to certain kinds of counterfactual claims. For, unlike the Indexicality answers, they deliver the desired results when it comes to counterfactuals like (2).

9.3.4 Language-Relativity and Metrics-Relativity

Next let's consider the two Relativity answers: if one adopts either of the Relativity answers, then the notion of lawhood is relativized – to a language (given Language-Relativity), a set of metrics (given Metrics-Relativity), or both.

It is commonly noted that our concept of lawhood is deeply related to modal notions, causal notions, and many other philosophical concepts.[34] If lawhood is a relation between a system and a language, or a system and a set of metrics, then this relativity will seep into these other notions as well. For example, although philosophers disagree on the details, it is generally thought that a vase cannot be fragile without certain counterfactuals about it being true, that one cannot perceive the breaking of the vase without bearing some kind of causal relationship to it, and that one cannot explain the breaking of the vase without describing some of the causal or nomic facts that led to its breaking. If lawhood is relative to a language or set of metrics, then it seems that the notions of fragility, perception, and explanation must all be language- or metrics-relative as well.

Thus, one worry about the Relativity answers is that the relativity of lawhood will spread to many of our other notions. And it seems prima facie implausible that nomic claims, causal claims, counterfactual claims, dispositional claims (such as claims about fragility), claims about what one is referring to (given a causal theory of reference), claims about psychological states (given a functionalist theory of mind), and so on, are language- or metrics-relative.

A more troubling worry is that this relativization leads to normative conflicts. (The following examples focus on Language-Relativity, though the same worries may be raised for Metrics-Relativity.) Suppose that agents ought to align their credences with what they believe the chances to be, in accordance with Lewis's Principal Principle (1980). If the laws determine the chances, and the laws are language-relative, then the chances are language-relative as well. But this yields conflicting

prescriptions about what we should believe. For instance, if I know that the laws relative to language L_1 assign a chance of 0.5 to a coin toss landing heads, and I know that the laws relative to language L_2 assign a chance of 0.6 to the same coin toss landing heads, then it seems that I am required to have a credence of both 0.5 and 0.6 in the coin toss landing heads.[35,36]

Another example: causal decision theorists maintain that, roughly, a prudentially rational agent will perform the act that she expects to causally bring about the most utility. But if the causal facts are language-relative, then how much utility an act causally brings about is also language-relative. And thus it seems that causal decision theory yields conflicting prescriptions regarding what we ought (prudentially) to do. For example, consider a scenario where the causal facts relative to language L_1 are such that smoking will cause cancer, while the causal facts relative to language L_2 are such that smoking is only statistically correlated with cancer. Assuming that one knows all of this, and that one finds smoking enjoyable, it seems that causal decision theory will prescribe both smoking and not smoking.

Similar worries arise for the normative prescriptions of moral theories like utilitarianism, which typically evaluate the net change in utility brought about by an act by evaluating the causal influence of that act. And legal prescriptions run into trouble as well. Suppose a company dumps toxins near a town's water supply, and the town experiences a spike in cancer rates. The company is legally obligated to pay compensation to the town iff the toxins it dumped nearby are the cause of this increase in cancer rates. Consider two languages, L_1 and L_2, such that the toxins cause-in-L_1 cancer but do not cause-in-L_2 cancer. Then the law yields conflicting prescriptions about whether the company owes the town any compensation.

Now, one might attempt to avoid these normative conflicts by relativizing norms as well. But this would require a wholesale revision of how we think about norms. And one might be wary of upending our conception of normativity in order to hold on to a particular account of laws.

9.3.5 Language-Salience

Finally, let's consider Language-Salience. According to this answer, candidate systems may be formulated in any language whatsoever. In order to rule out "trivial" candidate systems (for instance, a system consisting just of $\{\forall x \mathrm{F}x\}$, which is maximally informative and simple), we add a third desideratum: *salience*. A salient system is useful, unified, and explanatory. Because $\{\forall x \mathrm{F}x\}$ scores very low on salience, it cannot be the best system.

One might worry that the Language-Salience answer makes the laws of nature depend on us. For it is natural to think that assessments of salience – whether a system is useful, explanatory, etc. – implicitly rely on certain facts about we are like – what *we* find useful, explanatory, etc. If so, appealing to salience risks making the laws depend on particular features of our psychologies.

Note that this worry applies equally well to simplicity and informativeness. It may seem that assessments of simplicity and informativeness implicitly rely on certain facts about what we are like – what *we* find simple and informative. And if so, appealing to simplicity and informativeness risks making the laws depend on particular features of our psychologies.

But whether appealing to any of these desiderata makes the laws depend on us in this way hinges on how we answer the Metrics Question. Given Metrics-Objectivity, there are objective metrics of simplicity, informativeness, and salience that are independent of us – and so the laws are independent of us. Given Metrics-Indexicality, on the other hand, the metrics of simplicity, informativeness, and salience are those of the speaker – and so the laws depend on the speaker. So there is nothing about the Language-Salience answer, in and of itself, that threatens to make the laws depend on us. Language-Salience just inherits the "us-dependence" of whatever answer to the Metrics Question it is paired with.

A more pressing worry for Language-Salience is that the notion of salience is somewhat nebulous and underspecified. While one might raise the same complaint for simplicity and informativeness, it's plausible that salience is worse off in this regard. (Though see Hoefer 2007, 571–2, for reasons to think that simplicity and informativeness are not nearly as clear as generally supposed.[37]) For one might maintain that we have more of an intuitive grasp on the notions of simplicity and informativeness than we have on the notion of salience. And while proponents of Language-Salience have gestured towards ways of understanding the notion, it is reasonable to expect a more concrete characterization.

9.3.6 Taking Stock

All of the alternatives to the Objectivity answers to the Language and Metrics Questions face challenges of one kind or another. Given that we've put the Objectivity answers aside, which of the remaining answers are the most promising?

Begin with the Language Question. Language-Indexicality has trouble making room for disagreement, delivers counterintuitive results regarding counterfactuals, and, because the truth-values of nomic claims vary from speaker to speaker, makes the laws "us-dependent" in some sense. Language-Relativity leads to normative conflicts, and entails that the many notions intertwined with lawhood must be relativized as well. For these reasons, both Language-Indexicality and Language-Relativity strike us as unappealing.

Language-Rigidity has drawbacks as well, though they differ depending on the version we're considering. The idealized version of Language-Rigidity faces serious worries stemming from its appeal to an idealized scientific community. The actual version strikes us as more viable, but it too has uncomfortable results. First, there is the consequence that future scientists are likely to be mistaken about what the laws are. Second, there is the worry that the laws depend on what we're like in some sense. Still, the latter worry seems less pressing for Language-Rigidity than Language-Indexicality – for at least Language-Rigidity delivers the intuitively correct results regarding certain kinds of counterfactual claims.

Language-Salience seems to us to be the least problematic answer to the Language Question. The main worry facing this answer is that we do not have a fleshed out characterization of the notion of "salience" – and different characterizations of the notion may make this answer more or less plausible. Nonetheless, the task of providing a characterization seems feasible.

Next let's turn to the Metrics Question. Metrics-Indexicality and Metrics-Relativity have the same demerits as their Language counterparts, and so are unappealing for the same reasons. And, as before, the actual version of Metrics-Rigidity strikes us as more viable than the idealized version. But unlike Language-Rigidity, the worry that future scientists will be wrong about what the laws are seems less likely to arise for Metrics-Rigidity. All in all, this answer seems to be the most promising option.

The pair of answers that strike us as the most attractive – Language-Salience and Metrics-Rigidity – yield a Best System account similar to the one advocated by Loewer (2007) and Hoefer (2007). While one can raise worries for each of these answers, these worries seem to us to be less damaging than those facing the alternatives.

9.4 Extending the Account

In the previous section, we examined several ways of developing a Best System account of laws that does not require perfectly natural properties. We suggested that one promising route is to adopt Language-Salience and Metrics-Rigidity.

But natural properties play an important role in many Lewisian analyses – duplication, intrinsicality, supervenience, materialism, among others – in addition to Lewis's account of lawhood. Can these other analyses be modified so that they too do not require perfectly natural properties?

We suggest the following strategy: specify a set of properties that can effectively play the role of the perfectly natural properties in Lewisian analyses. Call the properties in this set the *surrogate* properties. Wherever a Lewisian analysis invokes perfectly natural properties, substitute these surrogates. For example, instead of taking individuals to be duplicates iff there is a bijection between their parts that preserves perfectly natural properties and relations, take them to be duplicates iff there is a bijection between their parts that preserves surrogate properties and relations.

The challenge is to specify the set of surrogates in some reductionist-friendly way. Our method begins with the account of laws presented above, which adopts Language-Salience and Metrics-Rigidity. On this account, the laws are given by the set of true sentences that, given our present metrics, best balances simplicity, informativeness, and salience. Our proposal is this: the surrogate properties are all and only the properties denoted by the predicates appearing in the best system.[38, 39, 40]

9.4.1 Three Worries

This proposal faces several challenges. Here we discuss three, in ascending order of importance.

One worry concerns the possibility of the world being very chaotic – so chaotic that there end up being few or no laws. If there are no laws, then there are no surrogate natural properties. If there are no surrogate natural properties, then none of Lewis's analyses are tenable.[41] So if there are no laws, then our proposal doesn't resurrect Lewis's analyses; it renders these analyses hopeless.

We don't take this worry to be compelling. For while it's true that our world *might* have had few or no laws, it's not true that our world *does* have few or no laws. Our proposal takes the predicates that appear in our actual laws to designate surrogate properties. As long as the actual world has laws, the possibility of worlds that do not is not a threat to the account.

A second worry is that the predicates that appear in the actual best system may denote properties that do not fit the roles Lewis envisioned. For instance, Lewis's analyses of duplication and intrinsicality entail that all the perfectly natural properties are intrinsic. Substituting surrogate properties for the perfectly natural ones, this entails that all the surrogate properties are intrinsic. Now consider the predicate "being five feet from an electron." If the system that best balances simplicity, informativeness, and salience employs this predicate, then this predicate denotes a surrogate property, and thus an intrinsic property. Intuitively, though, *being five feet from an electron* is not intrinsic.

There are two ways to resist this worry. First, it seems plausible that predicates of this sort will not appear in the actual best system. Given Language-Salience and Metrics-Rigidity, the best system is the one that does best at balancing our current, actual metrics of simplicity, informativeness, and salience. All else being equal, a system expressed in terms of predicates such as "being five feet from an electron," "being five feet from a proton," etc., is less simple than one that replaces these with a single relation, "being five feet from." Likewise, "being five feet from an electron" is less salient than "being five feet from"; all else being equal, a system employing the former predicate scores lower on salience than one employing the latter. Given the metrics we use to evaluate candidate systems, it seems unlikely that the system that comes out best will employ predicates that cause trouble in this way.

Second, one might say that if the best system does include monadic predicates of this sort, then we should bite the bullet. After all, our intuitions regarding intrinsicality are not sacrosanct. For example, we typically think of charge as an intrinsic property of particles. But if the system that best balances simplicity, informativeness, and salience takes the form of certain kinds of gauge theories,

then charge turns out to be extrinsic instead.[42] Insofar as the actual best system yields these kinds of unexpected results, we should take them to be discoveries, not counterexamples.

The third worry for this proposal concerns the intuitive possibility of properties that play the "naturalness" role but are not instantiated at the actual world. Call such properties *foreign properties*.

The proposal we've put forth has the naturalness role played by surrogate properties, and, necessarily, all surrogate properties are actual. As a result, this proposal rules out the possibility of foreign properties, and one might worry that this leads to counterintuitive results. For instance, this proposal entails that worlds alike with respect to their instantiation pattern of surrogates are duplicates. But if foreign properties are possible, then it seems that worlds alike with respect to their surrogates need not be duplicates – for one may instantiate a foreign property that the other does not.

Perhaps the best way to respond to this worry is to bite the bullet and deny the possibility of foreign properties.[43] There are a few reasons one might offer for thinking that this is not much of a cost.

First, there are independent reasons why one may want to reject the possibility of foreign properties. For instance, those who want to defend a combinatorialist theory of possibility (where possible worlds are "built up" out of actual world constituents) will reject foreign properties (e.g., see Armstrong 1989). Moreover, philosophers inclined towards the proposal we've sketched seem particularly likely to be suspicious of foreign properties. For those who dislike naturalness tend to dislike metaphysical weirdness in general, and a possible world where foreign properties are instantiated seems to fall into this category.

A second reason is that there are many possibilities concerning non-actual properties that this proposal *does* capture – and one might think that the demand that it also capture the possibility of foreign properties is too strong. For instance, this proposal does not rule out the possibility of alien properties (properties not instantiated at the actual world), for worlds with different instantiation patterns of surrogates may instantiate non-actual properties. Consider the property *being a golden mountain one mile high*. This property may very well be instantiated at some possible world, though it is alien to ours. Nor does this proposal rule out the possibility that the best system could have been different. There are possible worlds that differ from ours with respect to their instantiation patterns of surrogates, and many of these have best systems that differ from ours as well. This proposal does not even rule out the possibility that the predicates figuring in the best system could have been different. For, on this proposal, the best system of a world is the one that best balances informativeness, simplicity, and salience, evaluated using our actual metrics. This is compatible with there being a world where the instantiation pattern of surrogates is such that the system that best balances informativeness, simplicity, and salience is one whose predicates do not express surrogate properties. (Suppose, for instance, that the actual best system includes the predicate expressing the property *mass-density*. A world with a very different pattern of mass-density properties might be more saliently described using a predicate expressing *mass* instead, or even a predicate expressing some alien property.) Given this, one might claim that this proposal captures all the intuitions concerning non-actual properties that one can reasonably expect.

A third reason is that there are moves one can make that mitigate the counterintuitiveness of this result. For instance, following Armstrong (1989, 73), one might claim that although foreign properties are not *metaphysically* possible, they are *doxastically* possible. By distinguishing metaphysical from doxastic possibility, one can attempt to capture the intuition that there could have been foreign properties by claiming that the relevant notion of "could" is epistemic. (Of course, making this move incurs the obligation to provide an account of epistemic possibility; see Chalmers (2011) for one way of doing this.)

We have sketched a proposal for how to employ the kind of natural-property-free account of laws suggested by Loewer and Hoefer in the more ambitious project of doing without natural properties entirely. And as we've seen, several worries can be raised for this proposal. But we believe that there are viable ways in which a proponent of this proposal could respond.

9.5 Another Approach

In the previous section, we suggested one way to modify Lewis's accounts of similarity, duplication, etc., to avoid commitment to perfectly natural properties. We began by presenting a Best System account of lawhood that did not require perfect naturalness. We then took the properties expressed by the predicates appearing in the best system to be the "surrogate properties," and set those properties to play the role of the natural properties in Lewis's other accounts.

But there are other paths to selecting a set of surrogates. In this section, we briefly explore an alternative strategy. Instead of beginning with a Lewisian account of lawhood, this strategy begins with a Lewisian account of linguistic representation. We discuss a view presented by Williams (chapter 23, this volume), which modifies Lewis's account of linguistic representation to avoid appealing to primitive naturalness, and then sketch how one might employ this theory to select a set of surrogates to play the role of naturalness in other Lewisian analyses.

Let's begin with a rough sketch of Lewis's "Best Theory" account of linguistic representation. Consider the various ways a theory may assign semantic values to expressions. Some semantic theories line up with a community's linguistic practices better than others – these theories score highly on *fit*. Some semantic theories assign referents that are more eligible than others – these theories score highly on *eligibility*. Call the semantic theory that best balances fit and eligibility the "best theory."

The desiderata of fit and eligibility for semantic theories are analogous to the desiderata of informativeness and simplicity for theories of lawhood. A system of lawhood that scores highly on informativeness tells us a lot about what the world is like; a semantic theory that scores highly on fit tells us a lot about what the linguistic practices of a community are like. Similarly, a system of lawhood that scores highly on simplicity is one whose axioms are relatively short and uncomplicated; a semantic theory that scores highly on eligibility is one whose axioms – which state the semantic values assigned to lexical items – are relatively short and uncomplicated. (See Williams (2007 and chapter 23, this volume) for more on why the eligibility of semantic values assigned by the theory is equivalent to the simplicity of the theory's axioms.[44])

So we now have a basic template for Best Theory accounts of linguistic representation, analogous to the template for Best System accounts of lawhood. And, as before, this template is silent on two key questions. The first is the Language Question: what language should be used to formulate candidate semantic theories? Consider two theories that perfectly fit a community's patterns of use. The eligibility desideratum allows us to break the tie: the better theory is the one that assigns more eligible semantic values. But the degree to which a theory is eligible is just the degree to which its axioms are simple. And if candidate theories may be formulated in any language whatsoever, then the eligibility desideratum is too easy to satisfy – any theory can be made highly eligible, for the theory's axioms can be stated using a language that makes them as syntactically simple as you like.

The second is the Metrics Question: which metrics should we use to evaluate candidate semantic theories? Just as there are a number of ways to assess the simplicity and informativeness of a system, there are a number of ways to assess the fit and eligibility of a theory. And which theory comes out "best" will vary depending on the metrics we employ.[45]

One can answer these two questions in different ways. Lewis, for instance, offers the Metrics-Objectivity answer to the Metrics Question, according to which there are objectively correct evaluative metrics of fit, eligibility, and balance; and the Language-Objectivity answer to the Language Question, according to which the language in which to formulate candidate theories is the one whose predicates express perfectly natural properties.[46] In contrast, Williams (chapter 23, this volume) answers the Language Question in a way that avoids commitment to perfect naturalness. He offers what he calls a "parochial" version of the Best Theory account, which gives the Language-Rigidity answer to the Language Question: the language in which candidate semantic theories are formulated is a "refined, supplemented and improved" version of English, or "some interesting principled core (scientific English, or the final vocabulary of idealized science, perhaps)." (Williams follows Lewis in adopting Metrics-Objectivity.)

However one answers these two questions, the resulting semantic theory can be used to select a set of surrogate properties: formulate the semantic theory in the appropriate language, and let the surrogate properties be those expressed by the predicates appearing in its axioms.[47] So we now have two strategies for selecting a set of surrogate properties – a laws-based strategy and a semantics-based strategy. Which one should we pursue?

We do not think that one has to choose a single strategy for selecting surrogate properties. Once we eliminate primitive naturalness, it is no longer important that just *one* thing play all the roles Lewis had naturalness play.[48] Plausibly, the surrogates selected using the laws-based strategy are better suited to some of the naturalness roles, while the surrogates selected using the semantics-based strategy are better suited to others. For instance, one might think that the surrogates selected using the laws-based strategy are well suited to play the role of naturalness in analyses of duplication, intrinsicality, and related notions; while the surrogates selected using the semantics-based strategy are well suited to play the role of naturalness as reference magnets. Alternatively, one might combine the two strategies. One way is to take the set of surrogate properties to be the union of those selected using the laws-based strategy and the semantics-based strategy. Another way is to take candidate theories to be pairs of lawhood and semantic theories – a "laws-and-semantics" package – and then take the surrogate properties to be those that appear in the best laws-and-semantics pair. Which of these options is best is a question for further research.

9.6 Conclusion

The prospect of avoiding commitment to perfectly natural properties while holding on to a Lewisian analysis of lawhood is appealing to many philosophers. In this chapter we explored different ways one might pursue this project, and argued that a Best System account that adopts Language-Salience and Metrics-Rigidity, along the lines suggested by Loewer (2007) and Hoefer (2007), has the most hope of success. On this version of the Best System account of lawhood, the best system is formulated in the language of the candidate system that best balances simplicity, informativeness, and salience – where metrics of simplicity, informativeness, salience, and balance are our current, actual metrics. We then explored the possibility of modifying other Lewisian analyses, so that they too avoid commitment to a primitive distinction among properties. We suggested that the version of the Best System account developed in section 9.3 can be used to specify a set of "surrogate properties" – properties that can be substituted for perfectly natural properties in other Lewisian analyses.

Of course, there are several worries one might raise for this kind of proposal. But none of them seem insurmountable, and we've suggested some ways in which a proponent can respond. Given this, we take this proposal to be an interesting and viable way to maintain that there is no work for a theory of universals.

Acknowledgments

Many thanks to participants of the UC-Davis DEX Conference, the Midsummer Philosophy Workshop in Edinburgh, the Fundamentality Workshop in Birmingham, and audience members at New York University for helpful discussion. Thanks in particular to Elizabeth Barnes, Sara Bernstein, Phil Bricker, Tim Button, Louis deRosset, Sinan Dogramaci, Janice Dowell, Jesse Fitts, Shieva Kleinschmidt, Karen Lewis, Daniel Nolan, Zee Perry, Adam Sennet, Catherine Sutton, Lee Walters, and Alastair Wilson. Special thanks to John Roberts for extraordinarily extensive and helpful comments. Finally, this paper owes an especially large debt to Barry Loewer. As will be clear to anyone who has talked with Loewer about these issues, this paper is largely an exploration of his ideas.

Notes

1 Lewis restricts the laws to regularities entailed by the best system. But many of his successors have dropped this requirement, thus allowing laws constraining initial conditions (for example, see Loewer 2001; Hoefer 2007; Winsberg 2008).

2 Loewer writes: "I am very much attracted to [the Best System account] because of the way it incorporates the criteria physicists use for counting generalizations and equations as expressing laws and also because it, unlike many of its rivals, doesn't posit metaphysically primitive laws, primitive causal powers, propensities, governing relations, or other metaphysically heavy-duty and suspect entities" (2007, 313). Callender and Cohen echo the sentiment that the Best System account is appealing in part because it is not "metaphysically freighted":

> It is grounded in nothing more than the properties, individuals, and events in the world, and formal (deductive) relations defined over statements about these entities, that are already recognized by our best scientific descriptions of the way the world is. The modesty of [the Best System account's] extra-scientific apparatus has made the view seem attractive to thinkers who are inclined to defer to the best scientific descriptions of the world – both to Humeans (and others who forswear necessary connections) and more generally to those who prefer fewer metaphysical posits to more. (2009, 2–3)

3 For example, see Taylor 1993; Roberts 1999; Halpin 2003; Hoefer 2007; Loewer 2007; and Callender and Cohen 2009.

4 We follow Lewis ([1983] 1999, 41–2) in remaining neutral between understanding laws as propositions or interpreted sentences.

5 Languages may take many forms. Following Lewis, we restrict our attention to languages that take the form of something like second-order predicate logic. Moreover, these languages are *interpreted* languages. (If they weren't interpreted, one would not be able to evaluate the informativeness of sentences in the language.)

6 Lewis 1983.

7 See section 9.2.3 for more on how to understand the salience desideratum.

8 See Loewer 2007, 324–5.

9 Callender and Cohen's discussion suggests that they take Language-Relativity to provide an answer to the Metrics Question as well as the Language Question. For instance, they write:

> [We assess] the immanently strongest, simplest, and best balanced axiomatizations relative to a specific choice of basic kinds K (to a specific choice of basic predicates P_K). Given such an assessment relative to a choice of basic kinds K (predicates P_K), we can say that a true generalization is a law relative to K (P_K) just in case it appears in all the immanently Best Systems relative to the basic kinds K (basic predicates P_K). (2009, 21)

10 See Taylor 1993, 97.

11 See Halpin 2003, 151 and 156.

12 For passages that suggest the Indexicality answers, see pp. 142, 149, and 151. For passages that suggest the Relativity answers, see pp. 156 and 163–4.

13 However, Roberts does not commit himself to any version of the Best System account. Roberts (2009) provides a detailed defense of a different, though still Humean, account of laws.

14 "From the perspective of the aims of science the obvious trouble with '(x)Fx' is not that 'Fx' doesn't refer to a perfectly natural property but that '(x)Fx' is not a credible scientific theory. It is completely lacking in explanatory value. . . . The information in a theory needs to be extractable in a way that connects with the problems and matters that are of scientific interest" (Loewer 2007, 324–5).

15 "Here is a proposal. Let SL be a present language of science, say scientific English . . . A candidate for a final theory is evaluated with respect to, among the other virtues, the extent to which it is informative and explanatory about truths of scientific interest as formulated in SL or any language SL+ that may succeed SL in the rational development of the sciences" (Loewer 2007, 325).

16 "[P]art of what makes a truth a law is to an extent anthropomorphic since it involves human criteria of simplicity and informativeness and . . . explanatoriness and 'scientific interest.' But these may well be relatively objective notions" (Loewer 2007, 325). "It may turn out that the [Final Theory] entails that any intelligent creature that has a concept that plays a role like the concept of law plays for us will evaluate simplicity and informativeness in ways very similar to us" (327).

17 See Hoefer 2007, 571–2.

18

> The standards of simplicity, of strength, and of balance between them are to be those that guide us in assessing the credibility of rival hypotheses as to what the laws are. In a way, that makes lawhood depend on us – a feature of the approach that I do not at all welcome! But at least it does not follow that lawhood depends on us in the most straightforward way: namely, that if our standards were suitably different, then the laws would be different. For we can take our actual standards as fixed, and apply them in asking what the laws would be in various counterfactual situations, including counterfactual situations in which people have different standards – or in which there are no people at all. Likewise, it fortunately does not follow that the laws are different at other times and places where there live people with other standards. (Lewis [1980] 1986, 123)

19 "I used to think rigidification came to the rescue: in talking about what the laws would be if we changed our thinking, we use not our hypothetical new standards of simplicity and strength and balance, but rather our actual and present standards. But now I think that is a cosmetic remedy only. It doesn't make the problem go away, it only makes it harder to state" (Lewis [1994] 1999, 232).

20 One might wonder why Metrics-Rigidity is included among the answers that entail that lawhood is a "psychological matter." For it might seem that the laws are dependent upon one's psychology iff the following counterfactual is true: had one's psychology been appropriately different, then the laws would have been different. If so, then the Metrics-Rigidity answer would not make the laws dependent on one's psychology. Lewis, however, rejects this construal of psychological dependence, and maintains that there is a sense in which Metrics-Rigidity makes the laws dependent on our psychology (see Lewis 1994, 232).

21 There are "narrow" and "broad" versions of the Indexicality answers. On the narrow understanding, the language or metrics are indexed to the speaker. On the broad understanding, the language or metrics are indexed to the speaker's scientific community. For ease of discussion, we assume the narrow understanding in the text, although nothing hangs on this. Similarly, there are "actual" and "ideal" versions of the Indexicality answers. On the actual understanding, the language or metrics are indexed to the actual language or metrics of the party in question (either the speaker or the speaker's community). On the ideal understanding, the language or metrics are indexed to some idealized version of that party. In the section assessing the Rigidity answers, we argue against the idealized understanding of these answers, and the same arguments apply here as well. For ease of discussion, we assume the actual understanding in the text.

22 A somewhat related view takes "laws" to pick out the sentences entailed by the best system indexed to the language and metrics of the individual assessing a given utterance, rather than indexed to those of the speaker. We discuss this view in note (26).

23 Our discussion can be adapted to address Language-Indexicality in the following way. Let us say that Aristotle and Newton speak distinct but overlapping languages. The languages share terms such as "law," "language," etc. – whatever is needed so that they can intelligibly converse about laws, and make sense of sentences (1)–(5). But the languages differ with respect to their other predicates, in such a way that Aristotle's laws differ from Newton's. The worries raised in the text then apply to Language-Indexicality.

24 Thanks here to Sinan Dogramaci. For an interesting reply to these kinds of worries, see Roberts (2009, section 3.3) (though Roberts is evaluating these worries with respect to a somewhat different account of laws).

25 Do the Rigidity answers differ from a version of the Objectivity answers according to which the objectively correct language and metrics "happen" to line up with our own? It depends. If we allow ourselves a notion of "in virtue of" or "dependence," then we can articulate the difference between the two views in the following way: on the Rigidity answers, the right language and metrics to use depend on us; on the Objectivity answers, they do not (and the fact that the objectively correct language and metrics happen to line up with our own is just a coincidence). If we do not allow ourselves a notion of "in virtue of" or "dependence," then the two views seem to be equivalent. One can then understand our discussion as spelling out some worries for this particular version of the Objectivity answers (although the "arbitrariness" worry mentioned in note (30) may be more naturally construed as a "coincidence" worry).

26 One could understand the phrase "our language and metrics" in either a non-egalitarian way (there's some particular community, ours, whose language and metrics are used to determine the laws) or an egalitarian way (every community uses its own language and metrics to determine the laws). The view we discuss in the text adopts the non-egalitarian reading, since that's the view discussed in the literature. But the egalitarian reading is interesting enough to be worth a few remarks. On the egalitarian version of the Rigidity answers, the members of a community assess nomic claims using the language and metrics of their community; so the same utterance may be true when assessed by a member of one community and false when assessed by a member of another. Thus on this egalitarian view the truth values of nomic claims are relative to the assessor's community. This is an instance of the kind of Relativism described by MacFarlane (2005, 2009, 2013a, and 2013b). The egalitarian version of the Rigidity answers faces several worries. First, like the view discussed in the text, it seems likely that we will assess the nomic claims made by future scientists as false. (They won't be false when assessed by the future scientists, of course, but it is still counterintuitive that we should entertain these kinds of doubts about the nomic claims made by future scientists.) Second, since the same utterance can have different truth values when assessed by different people (see MacFarlane 2005), this view requires one to accept a kind of relativism about truth that some may find unappealing.

Finally, as we note in the section assessing the Relativity answers, lawhood is tied to many other notions (causation, dispositions, reference, mental states, etc.), and if laws are assessor-dependent, then these other notions are assessor-dependent as well. As a result, adopting assessor-dependence about lawhood implies that many more claims are assessor-dependent than advocates of relative truth generally believe. For recently proposed assessor-dependent views tend to be circumscribed to, say, epistemic modals, knowledge attributions, or claims about taste (see MacFarlane 2009, 2013a, and 2013b). If we adopt assessor-dependence about lawhood, then not only are nomic claims assessor-dependent, but so are causal claims, counterfactual claims, dispositional claims, claims about reference (given a causal theory of reference), claims about psychological states (given a functionalist theory of mind), and the like.

27 See also Carroll 1990, 201–2.

28 Loewer (2007) suggests something like this move. He characterizes the "final theory" (FT) in this way:

> FT maximally satisfies all the requirements that the tradition and practice of fundamental physics puts on a fundamental theory of the world. FT is true, simple, highly informative, comprehensive; FT reconciles relativity and quantum theory, explains statistical mechanical probabilities, and explains special science regularities, and so on. It does all this better than any alternative theory whether the alternative has ever been or ever will be thought up by anyone. There is no true theory that better than FT balances all these virtues. (2007, 322)

29 This is true on virtually every account of counterfactuals on offer; see Lewis (1979) and Maudlin (2007) for two examples at the opposite ends of the spectrum. (One also finds this close connection between laws and counterfactuals in Lange (2009), though he inverts the direction of dependence.)

30 Another way of developing this worry is to maintain that appealing to our language and metrics makes the laws seem "arbitrary" – why, out of all possible scientific communities, is *ours* the one whose language and metrics determine what the laws are? See Roberts 1999, S503.

31 "The worst problem about the best-system analysis is that when we ask where the standards of simplicity and informativeness come from, the answer may seem to be that they come from us. . . . I used to think rigidification came to the rescue . . . But now I think that is a cosmetic remedy only. It doesn't make the problem go away, it only makes it harder to state." (Lewis [1994] 1999, 232)

32 For example, given that our community's language is M and the corresponding laws are LAWS_{M}, the Rigidity answers entail that counterfactuals like the following are false:

(5) If our community's language were N, then the laws would not be LAWS_{M}.

And this seems to be the right result. Now, one might wonder whether the Rigidity answers run into trouble with respect to indicative conditionals. For example, the Rigidity answers allow for indicative conditionals like the following to be true:

(6) If our community's language is M, then the laws are LAWS_{M}, but if our community's language is N, then the laws are not LAWS_{M}.

But we do not think these indicative conditionals reveal a disturbing kind of "us-dependence," since indicative conditionals are largely epistemic (see Kratzer 1986). For example, suppose you know that the Objectivity answers to the Language and Metrics Questions are correct, and you know that the objectively correct language and metrics happen to line up exactly with those of your community. However, you don't know exactly what your community's language and metrics are. Then you could truly assert something like (6). But it doesn't follow that the Objectivity answers make the laws "us-dependent"! (Thanks here to Daniel Nolan and Louis deRosset.)

33 Here is one attempt. Suppose, as in note (25), we allow ourselves some notion of "in virtue of" or "dependence." Then one can spell out the worry in this way: the Rigidity answers entail that the laws of nature obtain in virtue of us having the language and metrics that we do, and that makes the laws unacceptably dependent upon us. But the notion of "in virtue of" is controversial, and not everyone will accept it: Lewis, for instance, claims that we do not have a "clear enough understanding of 'solely in virtue of' " (2001, 384). If the foremost proponent of perfect naturalness rejected this notion as too mysterious, then it's plausible that those who reject naturalness as too mysterious will as well. And if so, they will not be moved by this way of putting the worry.

34 See, for instance, Carroll 1994, 1–12.

35 This is an overly simplified description of the case; as described, there are still moves the proponent of relativization can make. For a more detailed and rigorous description of cases in which such conflicts arise, see Meacham 2014.

36 Yet another worry here is that, given a functionalist causal role account of belief, the beliefs one has are themselves relativized to a language. For instance, the causal facts given language L_1 might be such that one has a credence of 0.5 in P, while the causal facts given language L_2 might be such that one does not have a credence of 0.5 in P. Thus even if the chance of a coin landing heads is 0.5 relative to both L_1 and L_2, an agent who satisfies the Principal Principle relative to L_1 will fail to do so relative to L_2, since L_1 and L_2 assign her different credences.

37

> [C]loser inspection of Lewis's theory destroys the initial impression of tidiness. Simplicity and strength are meant to be timeless, objective notions unrelated to our species or our scientific history. But one suspects that if BSA advocates aim to have their account mesh with scientific practice, these notions will have to be rather pragmatically defined. Moreover, simplicity and strength are simply not clearly characterized by Lewis or his followers. We do not know whether initial conditions, giving the state of the world at a time (or a sub-region), should count as one proposition or as infinitely many (nor how to weigh the reduction of simplicity, whatever answer we give); we do not know whether deterministic laws are automatically as strong as can be, or whether instead some added chance-laws may increase strength at an acceptable price; if the latter, we do not know how to weigh the increase in strength so purchased. (Hoefer 2007, 571)

38 Lewis (1983) proposes that natural properties be used as "reference magnets," to help determine the referents of predicates and names. One might worry that replacing natural properties with surrogates in an account of reference raises circularity problems. The reference magnets are the properties referred to by the predicates in the actual best system; but we need reference magnets to determine what the referents of the predicates in the actual best system are!

The reason this is not a problem is that the languages in which candidate systems are expressed are *interpreted* formal languages, and the referents of their terms are stipulated from the outset (see note 5). So we don't need reference magnets to secure the referents of the predicates in the actual best system. (What Lewis is providing is a way to determine the referents of terms in *natural* languages. Here we do need to appeal to something like reference magnets, since the referents of terms in natural languages aren't fixed by construction.) (Thanks here to Elizabeth Barnes.)

39 One of the worries facing Lewis's account of laws is that several candidate systems might be tied for best, and so there might not be a unique best system. This worry might seem particularly pressing for the proposal we put forth, since the best system determines both the laws and the surrogate properties. One response to this worry is to maintain that, in the case of ties, the intersection of the properties referred to in the best systems provide the surrogates. A second response is to take the union of these properties to be the surrogates. A third response is to note that the mere *possibility* of ties is not a problem for this proposal, since the *actual* best system is used to select the surrogate properties. And, as Lewis says, "we haven't the slightest reason to think the case [of ties] really arises" ([1994] 1999, 233).

40 And for quantitative properties (like *three grams mass*), one might want to say that for if any quantitative property appears in the best system, then all quantitative properties belonging to the same family (e.g. *mass*), count as surrogates.

41 Take, for instance, Lewis's analysis of duplication: two things are duplicates iff there's a bijection between their parts that preserves the perfectly natural properties and relations. If there are no surrogate natural properties, then all objects are alike with respect to their surrogates – they all have none. Together with Lewis's analysis of duplication, this entails that any mereologically isomorphic objects are duplicates. But that's clearly the wrong result. Moreover, this result will filter into any account that makes use of the notion of duplication. Consider Lewis's analysis of intrinsicality: a property is intrinsic iff it never divides duplicates. If all mereologically isomorphic objects are duplicates, then nearly every property is one that divides duplicates, and so nearly every property is extrinsic. But that, too, is the wrong result.

42 See Maudlin 2007, 78–103; and Arntzenius 2012, 183–93.

43 If one believes that modal facts are conventional or constructed in some sense, then there may be a way to accommodate the possibility of foreign properties. Begin with a linguistic ersatzist theory, such as the one offered by Joseph Melia (2001 and 2003, 165–72). Very roughly, linguistic ersatzism is the view that possible worlds are maximal consistent sets of sentences. Let the language used to formulate the sentences have a predicate for every actual property and a name for every actual individual. Melia suggests that this language be enriched with *pseudo-predicates* – expressions that function just like predicates but do not refer to any actual property – which can be used to represent non-actual properties. One might build on Melia's proposal by introducing a second-order pseudo-predicate, "is a foreign property," and stipulating that this pseudo-predicate satisfies certain modal constraints: (a) it applies to no actual properties, (b) it applies to some proper subset of pseudo-predicates, and (c) it applies to the same pseudo-predicates at every world. One could then introduce a disjunctive second order predicate, "is foreign or appears in the actual best system," and use this predicate to play the role of naturalness predicate in Lewis's accounts. (Similarly, one might accommodate the possibility of foreign properties by adopting a form of modal fictionalism; see Rosen 1990.)

44

> [W]e could assign a 'degree of eligibility' to each semantic value for a lexical item featuring in the semantic theory – a measure that reflects the syntactic complexity of a clause that assigns that semantic value to an expression. For example, the degree of eligibility of the property *being human* gives a measure of how much syntactic complexity is added to semantic theory by a clause assigning that property to the predicate 'is human'. The overall eligibility of a theory is thus just another way of measuring the syntactic complexity of that theory *when spelled out in primitive terms*. Measuring simplicity of a

> theory by its syntactic complexity when spelled out in primitive terms, or by overall eligibility of the semantic values assigned, are thus one and the same thing. (Williams 2007, 375–6)

45 As has been frequently noted, characterizing simplicity purely in terms of syntactic complexity, together with the assumption that the predicates that figure in the semantic theory also figure in fundamental physical theories, yields implausible results. See Schaffer 2004; Hawthorne 2006, 205–6; and Williams, chapter 23, this volume.

46 See Lewis (1975) for remarks on fit, ([1984] 1999, 65–6) for remarks on eligibility, and ([1983] 1999, 47–8 and 51–4) for remarks on both fit and eligibility. See also Williams, chapter 23, this volume.

47 How well these surrogates fit into Lewis's accounts will depend on how one answers the Language and Metrics Questions. While we are not here tackling the issue of which answers are most promising for the best theory account of linguistic representation, much of our discussion in section 9.3 applies here as well.

48 For an in-depth discussion of the (in)compatibility of the many roles Lewis has naturalness play, see Dorr and Hawthorne 2014.

References

Armstrong, D.M. (1989). *A combinatorial theory of possibility*. Cambridge: Cambridge University Press.

Arntzenius, F. (2012). *Space, time and stuff*. Oxford: Oxford University Press.

Callender, C. & Cohen, J. (2009). A better best system account of lawhood. *Philosophical Studies*, 145(1), 1–34.

Carroll, J. (1990). The Humean tradition. *Philosophical Review*, 99(2), 185–219.

Carroll, J. (1994). *Laws of nature*. Cambridge: Cambridge University Press.

Chalmers, D. (2011). The nature of epistemic space. In A. Egan & B. Weatherson (Eds.), *Epistemic modality*. Oxford: Oxford University Press.

Dorr, C. & Hawthorne, J. (2014). Naturalness. *Oxford studies in metaphysics, Vol. 8*. Oxford: Oxford University Press.

Halpin, J.F. (2003). Scientific law: a perspectival account. *Erkenntnis*, 58(2), 137–68.

Hawthorne, J. (2006). Epistemicism and semantic plasticity. *Oxford studies in metaphysics, Vol. 2*. Reprinted in J. Hawthorne (2006), *Metaphysical Essays*, pp. 185–210. Oxford: Oxford University Press.

Hoefer, C. (2007). The third way on objective probability: a skeptic's guide to objective chance. *Mind*, 116, 549–96.

Kratzer, A. (1986). Conditionals. *Chicago Linguistics Society*, 22(2), 1–15.

Lange, M. (2009). *Laws and lawmakers: science, metaphysics, and the laws of nature*. Oxford: Oxford University Press.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, 13(4), 455–76.

Lewis, D. (1980). A subjectivist's guide to objective chance. In R. Jeffery (Ed.), *Studies in inductive logic and probability*, Vol. II. Oakland: University of California Press. Reprinted with postscripts in Lewis (1986), pp. 83–132.

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, 61, 343–77. Reprinted in Lewis (1999), pp. 8–55.

Lewis, D. (1984). Putnam's paradox. *Australasian Journal of Philosophy*, 62, 221–36. Reprinted in Lewis (1999), pp. 56–77.

Lewis, D. (1986). *Philosophical papers*, Vol. II. Oxford: Oxford University Press.

Lewis, D. (1994). Humean supervenience debugged. *Mind*, 103, 473–90. Reprinted with postscripts in Lewis (1999), pp. 224–47.

Lewis, D. (1999). *Papers in metaphysics and epistemology*. Cambridge: Cambridge University Press.

Lewis. D. 2001. Redefining "Intrinsic". *Philosophy and Phenomenological Research*, 63, 381–98.

Loewer, B. (2001). Determinism and chance. *Studies in the History of Modern Physics*, 32, 609–20.

Loewer, B. (2007). Laws and natural properties. *Philosophical Topics*, 35(1/2), 313–28.

MacFarlane, J. (2005). Making sense of relative truth. *Proceedings of the Aristotelian Society*, 105(3), 321–39.

MacFarlane, J. (2009). Epistemic modals are assessment-sensitive. In A. Egan & B. Weatherson (Eds.), *Epistemic modality*. Oxford: Oxford University Press.
MacFarlane, J. (2013a). Assessment sensitivity: relative truth and its applications. Unpublished draft.
MacFarlane, J. (2013b). Relativism and knowledge attributions. Unpublished draft.
Maudlin, T. (2007). *The metaphysics within physics*. Oxford: Oxford University Press.
Meacham, C.J.G. (2014). Chance and the conflicts problem. In A. Wilson (Ed.), *Chance and temporal asymmetry*. Oxford: Oxford University Press.
Melia, J. (2001). Reducing possibilities to language. *Analysis*, 61(1), 19–29.
Melia, J. (2003). *Modality*. Montreal: McGill-Queen's University Press.
Roberts, J. (1999). "Laws of nature" as an indexical term: a reinterpretation of Lewis's best-system analysis. *Philosophy of Science*, 66(3), S502–S511.
Roberts, J. (2009). *The law-governed universe*. Oxford: Oxford University Press.
Rosen, G. (1990). Modal fictionalism. *Mind*, 99(395), 327–54.
Schaffer, J. (2004). Two conceptions of sparse properties. *Pacific Philosophical Quarterly*, 85(1), 92–102.
Taylor, B. (1993). On natural properties in metaphysics. *Mind*, 102(405), 81–100.
Williams, J.R.G. (2007). Eligibility and inscrutability. *Philosophical Review*, 108(3), 361–99.
Winsberg, E. (2008). Laws and chances in statistical mechanics. *Studies in History and Philosophy of Modern Physics*, 39, 872–88.

10

Hume's Dictum and Metaphysical Modality

Lewis's Combinatorialism

JESSICA WILSON

Introduction

Many contemporary philosophers accept a strong generalization of Hume's denial of necessary causal connections, in the form of Hume's dictum (HD), according to which there are no metaphysically necessary connections between distinct, intrinsically typed entities.[1] Such widespread endorsement of HD is somewhat puzzling, however, since proponents typically do not accept Hume's empiricism; nor (as I have argued elsewhere[2]) is HD motivated either as analytic, as synthetic *a priori* (motivated by intuitions we have no good reason to question), or as presupposed by the best account of counterfactuals. Tacit in David Lewis's work, however, is a promising potential motivation for HD, according to which one should accept HD as presupposed by the best account of the range of metaphysical possibilities – namely, a combinatorial account, applied to spatiotemporal fundamenta. Here I elucidate and assess this Ludovician motivation for HD. In section 10.1 I refine HD, and note its key, recurrent role in Lewis's work, as reflected in his thesis of Humean supervenience and his accounts of laws of nature and counterfactuals. In section 10.2 I present Lewis's specific appeal to HD as providing a broadly axiomatic generating basis for the space of metaphysical modality, and canvas the prima facie advantages of the resulting combinatorial principle – HD (L-combinatorialism) – as being *principled, extensionally sufficient* (in particular, leaving "no gaps" in the space), and *modally reductive*. Most criticisms of Lewis's combinatorialism have targeted seeming ways in which the theory overgenerates the desired space, letting in as possible what, by some or other lights, is impossible. In section 10.3 I rather argue that HD (L-combinatorialism) seriously *undergenerates* the desired space, for possibilities involving broadly scientific entities in particular, in three different ways. For each way I argue that available means of overcoming the undergeneration either fail to close the gap, undermine the claim that HD (L-combinatorialism) is a principled generator of metaphysical modal space, undermine the reductive status of Lewis's combinatorialism, or call into question the truth of HD. These results don't entirely close off a modal combinatorial motivation for HD, however; as I discuss in section 10.4, there are other combinatorial accounts on offer which presuppose HD, and more generally the end game here depends on whether any comparably principled alternative account

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

of the space of metaphysical modality can do better. Moreover, discussion of the foundational role HD plays in Lewis's philosophy suggests another potential motivation for HD – namely, as presupposed by the best overall "systematic" philosophy. As in so many areas of philosophy, Lewis's views, both general and specific, remain the ones to beat.

10.1 HD and Its Recurrent Role in Lewis's Work

10.1.1 From Hume to HD

Hume's version of his dictum occurs during his investigation into the source of the idea of causal connection. Recall Hume's methodology in *A Treatise of Human Nature*:

> To begin regularly, we must consider the idea of causation, and see from what origin it is deriv'd. 'Tis impossible to reason justly, without understanding perfectly the idea concerning which we reason; and 'tis impossible perfectly to understand any idea, without tracing it up to its origin, and examining that primary impression, from which it arises . . . Let us therefore cast our eye on any two objects, which we call cause and effect, and turn them on all sides, in order to find that impression, which produces an idea of such prodigious consequence. (1978 [1739], Part III, S II)

After arguing that in experience of singular instances of casual relation we find no impression of necessary efficacy, Hume considers and rejects the suggestion that ideas of necessary causal connections might arise from broadly demonstrative inference, in what is the most explicit statement of Hume's version of his dictum:

> There is no object, which implies the existence of any other if we consider these objects in themselves, and never look beyond the ideas which we form of them. Such an inference wou'd amount to knowledge, and wou'd imply the absolute contradiction and impossibility of conceiving any thing different. But . . . 'tis evident there can be no impossibility of that kind. (Part III, S VI)

It is worth pointing out that this line of thought makes good sense if one is an empiricist of Hume's comparatively strict variety. As we learned at mother's knee, on Hume's view our ideas are ultimately grounded in "simple" sense impressions, with more complex ideas and associated beliefs being built up using a highly restricted set of associative elements, which includes only resemblance, spatiotemporal contiguity, and causation (and where the last turns out to be a construction of the first two). Most importantly, Hume's framework rejects inference to the best explanation (IBE) – for example, to the existence of unobservable forces, powers or dispositions necessarily connecting causes and effects – as a warranted mode of inference. Barring perception of modality, the only way such a framework can accommodate necessary connections between distinct existences is if these hold as a matter of meaning or definition, accessible to reason; so far as broadly scientific goings-on are concerned, we must rest with the fairly superficial appearances or (resemblance- or continguity-based) constructions thereof; and connections between these are plausibly always contingent. There is, to expand on Hume's favored example, no contradiction in supposing that the superficial form of one billiard ball might interact in some unexpected way with the superficial form of another billiard ball.

The contemporary version of Hume's dictum, HD, is both more general and more specific than Hume's version. In HD, Hume's talk of objects and events is generalized to talk of effectively any goings-on whatsoever, including properties, states of affairs, facts, or other varieties of being. His talk of implication is generalized as talk of necessity, and more specifically of metaphysical necessity. His talk of entities "considered in themselves" is made precise as talk of entities characterized, or typed,

in terms of their intrinsic features. And his indirect talk of entities that are distinct ("any other") is qualified, at least typically, as requiring not just bare or numerical distinctness, but rather "whole" distinctness, understood, for example, in terms of failure of spatiotemporal overlap. The general contemporary version of Hume's dictum is then as follows:

> **HD:** There are no metaphysically necessary connections between wholly distinct, intrinsically typed, entities.[3]

Such generalizations and specifications make for a more interesting philosophical thesis, but insofar as HD is supposed to be so generally applicable it requires correspondingly greater motivation. It is here that contemporary support of HD poses something of a puzzle. To start, contemporary proponents of HD do not accept Hume's strict empiricism, and as such cannot cite his reasons for endorsing even his restricted version of the thesis. Moreover, perhaps the most salient reason for contemporary rejection of strict empiricism lies in contemporary acceptance – shared by proponents of HD – of IBE as a warranted mode of inference. If proponents of HD are comfortable with IBE, what motivates their thinking that we cannot or should not so infer to the existence of (perhaps unobservable) necessary connections between wholly distinct, intrinsically characterized, entities? Hence it is that, notwithstanding widespread acceptance of HD, there remains a serious question about what, post-empiricism, motivates this acceptance.

10.1.2 HD's Recurrent Role in Lewis's Work

As prefigured, I will later consider whether post-empiricist acceptance of HD can be taken to lie in its serving as an appropriately principled generator of the space of metaphysical modality. But we can gain some antecedent insight into the shared motivation(s) for contemporary and historical versions of HD by considering the foundational role HD plays in other of Lewis's projects.

10.1.2.1 Humean supervenience (HS) Lewis says, in introducing the second volume of his collected papers, that "many of the papers, here and in Volume I, seem to me in hindsight to fall into place within a prolonged campaign on behalf of the thesis I call 'Humean supervenience'" (1987a, ix). He then characterizes this thesis as follows:

> Humean supervenience is named in honor of the greater denier of necessary connections. It is the doctrine that all there is to the world is a vast mosaic of local matters of particular fact, just one little thing and then another. . . . We have geometry: a system of external relations of spatiotemporal distance between points. Maybe points of spacetime itself, maybe point-sized bits of matter or aether or fields, maybe both. And at those points we have local qualities: perfectly natural intrinsic properties which need nothing bigger than a point at which to be instantiated. For short: we have an arrangement of qualities. And that is all. There is no difference without difference in the arrangement of qualities. All else supervenes on that. (ix)

It was ultimately in service of this broader project that Lewis was concerned to establish that laws of nature, counterfactuals, causation, mental states, and other phenomena supervene on the distribution of fundamental intrinsic qualities. Now, why, exactly, does Lewis name his doctrine in honor of Hume, and more specifically in honor of Hume qua "great denier of necessary connections"? First note that the question here isn't motivated by the fact that (notwithstanding Lewis's use of the indicative in characterizing the supervenience at issue) supervenience theses typically record necessary connections between distinct entities; for after all the supervenient and base entities here aren't

wholly distinct. Rather, the question is motivated by the fact that Lewis departs from Hume in accepting that IBE might support the posit of entities beyond the reach of (at least in being underdetermined by) sensory experience – for example, the posit of the fundamental physical qualities (that is: properties) that Lewis takes to enter into the supervenience base. As such, what guarantees that Humean supervenience is genuinely "Humean"? In particular, what prevents wholly distinct occupants of the mosaic from being necessarily connected?

The question becomes more pressing upon recognizing that Lewis's specification of the fundamental properties as "intrinsic" doesn't suffice to rule out this non-Humean possibility, though he often speaks as if it does. For a property may be intrinsic, in that its instantiation does not require the existence or instantiation of any other objects or properties (in evocative terms: it could be instanced by an object in a "lonely" world), and yet still be "modally loaded," in that, for example, the property is necessarily such that when instanced *in certain circumstances*, it (its instance) brings about certain effects. What then prevents the "intrinsic" fundamental physical properties in Lewis's supervenience base from standing in various necessary connections to other (actually or possibly instanced) properties?

The answer to this question is, I think, that Lewis is simply *stipulating* that the local, intrinsic fundamental properties satisfy HD; alternatively, perhaps he thinks science as it stands gives us reason to think that fundamental physical properties are not modally loaded (though what would count for or against this claim is unclear). Either way, for clarity's sake it would be better to use an expression such as "Hume-intrinsic" rather than just "intrinsic" in characterizing Lewis's supervenience base, to flag that the elements of the base are assumed not to be irreducibly modal. More importantly, here we have arrived at the reason why, in spite of its departures from various aspects of Hume's empiricism, Lewis's supervenience thesis is appropriately deemed Humean; namely, Lewis is on Hume's side so far as the status of necessary connections is concerned. Hence in describing the sort of world he has in mind as involving "just one little thing and then another" Lewis echoes Hume's famous remark:

> [U]pon the whole, there appears not, throughout all nature, any one instance of connexion which is conceivable by us. All events seem entirely loose and separate. One event follows another; but we never can observe any tie between them. They seem conjoined, but never connected. (1978 [1739], Part II)

Lewis agrees; and this agreement is registered in his assumption that the fundamental qualities in the Humean supervenience base conform not just to physicalism, but to HD.

10.1.2.2 Lewis's account of laws of nature Given that the elements in the Humean supervenience base are to conform to HD, Lewis requires an account of laws of nature on which such laws are contingent, such that elements in the base might enter into causal or other nomological connections different from those into which they actually enter. Hence it is that Lewis says that he is "prepared to take the offensive against alleged unHumean lawmakers," including Armstrong's account of laws of nature as involving second-order relations of so-called "necessitation" between universals, and accounts on which laws are grounded in modally loaded dispositions or powers, as per Swoyer (1982), Shoemaker (1998), and Bird (2007). Hume, of course, grounded (causal) laws of nature in experienced regularities. Even objectively construed, however, a regularity account faces well-known difficulties, both in distinguishing accidental from genuine causal regularities, and in accommodating probabilistic laws and laws that are singly or even never instanced. One of Lewis's notable contributions is in developing a more sophisticated account of laws as grounded in spatiotemporal patterns of events, associated with his "Best System" theory of laws (see Lewis 1973, 1987a, and 1994). Lewis summarizes:

> Few would deny that laws of nature, whatever else they may be, are at least exceptionless regularities. Not all regularities are laws, of course. But, following the lead of (a short temporal segment of) Ramsey, I suggest that the laws are the ones that buy into those systems of truths that achieve an unexcelled combination of simplicity and strength. That serves the Humean cause. For what it is to be simple and strong is safely noncontingent; and what regularities there are, or more generally, what candidate systems of truths, seems to supervene safely on the arrangement of qualities". (1987a, xi)

As with Hume's original conception, an account of laws as the best systematization of the spatiotemporal arrangement of Hume-intrinsic properties accommodates the supposed contingency of laws, for while different arrangements of such properties might be associated with the same set of laws, in general different arrangements will be differently systematized. Hence Lewis's Best System account of laws conforms to the presupposition of HD, as applying, in particular, to the elements of the Humean mosaic.

10.1.2.3 Lewis's account of counterfactuals With spatiotemporal arrangements of Hume-intrinsic properties and Best-System-theoretic laws of nature in hand, we can now consider the role HD plays in Lewis's preferred account of counterfactuals – that is, of subjunctive conditionals whose antecedents may be false, schematically along lines of 'P > Q' (if it were/had been that P, it would be/would have been that Q). The approach to counterfactuals Lewis favors is a similarity-based possible worlds account (see Stalnaker 1968; Lewis 1973). Roughly (sufficient for present purposes), and granting truth in cases of vacuity, such an account runs as follows:

> P > Q iff (i) there are no possible P-worlds, or (ii) some P&Q world is comparatively more similar overall ("closer") to the actual world than any P&-Q world.[4]

What does comparative overall similarity come to? Though Lewis initially took this to be primitive (1973, 75–7), in his (1979) he refined the relevant understanding of similarity to one involving four, differently weighted respects:

(1) It is of the first importance to avoid big, widespread, diverse violations of law
(2) It is of the second importance to maximize spatiotemporal region of perfect match of particular fact
(3) It is of the third importance to avoid even small, localized, simple violations of law
(4) It is of little or no importance to secure approximate similarity of particular fact (1979, 472)

Lewis endorses these respects, so ordered, as needed in order to preserve *CF asymmetry*, according to which future but not past states counterfactually depend on present states:

> The way the future is depends counterfactually on the way the present is. If the present were different, the future would be different . . . Not so in reverse. Seldom, if ever, can we find a clearly true counterfactual about how the past would be different if the present were somehow different. (1979, 32)

So, for example, consider the following counterfactual, assumed to have an actually false antecedent, under the supposition that the actual laws are deterministic:

> If these two electrons were next to one another, they would repel each other.

The 'P-worlds' – where the electrons are next to each other – most similar by lights of Lewis's weighting will turn out to be worlds where electrons enter into different laws than they actually do. Why

so? To start, in worlds where the laws are the same as the actual laws, the initial conditions must be changed in order to implement the antecedent; hence antecedent and actual pasts completely differ, making for large dissimilarity in matters of particular fact (against a desideratum of the second importance). By way of contrast, in some worlds with different laws, the actual and counterfactual past exactly match until just before the antecedent event occurs, at which time there are a few local violations of law of the sort needed to implement the antecedent. Since avoiding small, local violations of law is only of the third importance in the weighting, such a "past-fixing" P-world is more similar to the actual world than a "law-fixing" P-world. What about the future of the closest P-worlds? Does it similarly follow from the weighting that the closest P-worlds are ones where a minor variation from the actual laws somehow undoes whatever events brought the electrons together? No, says Lewis, for undoing all the propagating traces of these events requires many departures from actual law, adding up to a big, widespread, violation of law. Since avoiding such big violations of law is of the first importance, and since preserving approximate similarity of particular fact (here, as regards future events) counts for little or nothing, the future in the closest P-world will be one where events unfold just as they would were the actual laws in place. The electrons will repel, and the counterfactual will turn out both true and compatible with *CF asymmetry*, as desired.

More generally, Lewis's strategy for accommodating *CF asymmetry* relies on HD (causal) – HD as applied to the case of causal/nomological connections – since implementing this strategy requires that entities of the kind that actually exist may enter into different laws. Schaffer summarizes:

> [Lewis's] account of counterfactuals requires miracles (slight variations of the actual laws) in order to implement their antecedents. That is, to implement the antecedent that there are like charges at a given location (assuming this to be actually false), we need to imagine some miraculous swerving of say, two electrons, that brings them to said location. Assuming that the actual laws are deterministic . . . such a miraculous swerving [of electrons] will require a slight violation of the actual laws. Hence the laws of the nearest possible world in which there are like charges here must be just slightly different from the actual laws. Thus to implement the counterfactual antecedent, one needs worlds with actual properties but alien laws. This is contingentism. (2009, 216)

Schaffer (2009) suggests that this connection, along with the superiority of Lewis's account of CFs, provides support for contingentism/HD (causal). In Wilson (forthcoming) I argue that this IBE doesn't go through.[5] Here I simply want to flag the crucial role that HD plays in Lewis's account of counterfactuals.

10.2 HD and Lewis's Combinatorialism

As above, HD is foundational in nearly every aspect of Lewis's framework. The most powerful role that HD plays in Lewis's system, however, concerns its providing a basis for, as Lewis puts it, a "principle of plentitude" that will guarantee "that the worlds are abundant, and logical space is somehow complete" (1987a, 86). It is with respect to this role, I believe, that (an instance of) the most promising IBE for HD is to be found.[6] I will start by presenting Lewis's motivations for appealing to HD as a broadly axiomatic combinatorial generator of the space of possibility.[7] Some of these specifically pertain to Lewis's concrete modal realism, according to which possible worlds are of the same (concrete, particular) ontological type as our very own actual world, with combinatorial elements correspondingly taken to be spatiotemporal fundamenta or Hume-intrinsic duplicates thereof. As I will next try to bring out, however, the deeper motivation for Lewis's appeal to HD is to some extent neutral both on the metaphysics of possible worlds and on the preferred base of entities suitable for recombination.

10.2.1 HD as a Principle of Recombination

To see the deeper motivation for HD, it's useful to start by considering how the space of *syntactically* logical possibility is generated. On this conception, while 'There is something that is a bachelor and is not a bachelor', and 'There is something that is both red and not red' are each impossible, 'There is a married bachelor' and 'There is something that is both red and green all over' are each possible. Here, the job of generating the precise boundaries of the space of broadly syntactic possibility is done by a principle of syntactic consistency, according to which a represented state of affairs is syntactically logically possible iff it is syntactically consistent.

Though the space of syntactically logical possibility admits of consistency as a single principled generator, the associated conception of possibility is too weak to be of interesting use in characterizing what is metaphysically possible. It is metaphysical or "broadly" logical possibility that is of interest to metaphysicians. As Sider puts it:

> There are . . . different "strengths" of necessity and possibility, which can be signified by modal words (like 'can') in different contexts. Philosophers have tended to concentrate on a very broad sort, so-called "metaphysical" possibility and necessity. . . . What is not metaphysically possible? Almost everyone agrees that contradictions are metaphysically impossible – it is metaphysically impossible to both give a talk in California and also not to give a talk in California. And everyone who accepts the legitimacy of the notion of analyticity – of truth that is in some sense guaranteed by meaning – agrees that the negations of analytic sentences like 'all bachelors are unmarried' are impossible. But it is usually thought that there exist further impossibilities. Examples might include the existence of a round square, someone's being taller than himself, someone's being in two places at once, George W. Bush's being a donkey, there existing no numbers, and there existing some water that is not made up of H_2O. (2003, 181)

But how is the space of metaphysical modality to be generated? Is there any principle that can serve as the basis of metaphysical/broadly logical possibility in the non-arbitrary, extensionally adequate way that the principle of consistency does vis-à-vis the space of syntactically logical possibility? The appeal to HD in Lewis's and other combinatorial theories of modality is, I suggest, best seen as motivated by its promise in providing such a non-arbitrary, extensionally correct generator of metaphysical modal space.

That Lewis intends HD to serve as an extensionally correct generator is indicated by his concern with ensuring that the space of (broadly) logical possibility is "complete," without "gaps"; that he intends HD to serve as a principled or systematic such generator is indirectly suggested by his consideration of an alternative two-part principle as doing this work, according to which:

(1) absolutely every way that a world could possibly be is a way that some world is, and
(2) absolutely every way that a part of the world could possibly be is a way that some part of some world is (1986, 86)

Lewis's first concern with this two-part principle is that, if worlds are understood as concrete entities, then the principle ends up being contentless:[8]

> [G]iven modal realism, it becomes advantageous to identify 'ways a world could possibly be' with worlds themselves. Why distinguish two closely corresponding entities: a world, and also the maximally specific way that world is? Economy dictates identifying the 'ways' with worlds. (1986, 86)

But then (substituting) that makes (1) and (2) contentless:

(1) absolutely every world is a world, and
(2) absolutely every world that is part of a world is a world that is part of some world (perhaps itself).

As Lewis notes, one can avoid the triviality results by reading the two-part principle in epistemic terms, so that (1) says that every way we think a world could possibly be is a way that some world is; but in that case (1) "indiscriminately endorses offhand opinion about what is possible" (1986, 87). What is wanted is a principled and accurate *metaphysical* guide to the space of metaphysical possibility.

It is at this point that Lewis famously invokes a version of HD as his preferred generator of metaphysical modal space:

> We need a new way to say . . . that there are possibilities enough, and no gaps in logical space. To which end, I suggest that we look to the Humean denial of necessary connections between distinct existences. To express the plentitude of possible worlds, I require a *principle of recombination* according to which patching together parts of different possible worlds yields another possible world. Roughly speaking, the principle is that anything can coexist with anything else, at least provided they occupy distinct spatiotemporal positions. Likewise, anything can fail to coexist with anything else. (Lewis 1986, 87–8)

Various qualifications are encoded in or ensue from Lewis's appeal to HD qua combinatorial principle.

First, the combinatorial elements to which HD is intended to apply are occupants of space–time. Following a common understanding of what it is for spatiotemporally located entities to be "wholly distinct," HD is guaranteed to apply only to non-overlapping such entities ("at least provided they occupy distinct spatiotemporal positions," Lewis 1986, 88).

Second, the combinatorial elements may be either actual or (merely) possible. (Given Lewis's concrete modal realism, the reference to possibilia here is supposed not to invoke irreducible circularity, for reasons I discuss in section 10.3.) Hence, as Lewis observes, if there could be a dragon and there could be a unicorn (neither of which presumably actually exist), then the principle would allow for the possibility of a dragon and a unicorn existing side by side. A minor complication here is that in discussing alien worlds, involving alien individuals or (more importantly) alien properties, Lewis says "it won't do to say that all worlds are generated by recombination from parts of this world, individuals which are possible because they are actual. We can't get the alien possibilities just by rearranging non-alien ones. Thus our principle of recombination falls short of capturing all the plenitude of possibilities" (1986, 92). This makes it sound as if Lewis *doesn't* take his recombinatorial principle to be a complete generator of metaphysical modal space. But since there was no presupposition that the principle applied only to actual spatiotemporal occupants, the principle doesn't *really* (in this respect) fall short, and indeed Lewis goes on to note that "Although recombination will not generate alien worlds out of the parts of this world, it nevertheless applies to alien worlds. . . . Anything alien can coexist, or fail to coexist, with anything else alien, or with anything else not alien, in any arrangement permitted by shape and size" (92). It is the more general recombinatorial principle that is ultimately at issue in Lewis's combinatorialism.

Third, reflecting Lewis's concrete modal realism, and associated supposition that possible worlds do not "overlap", applications of the principle involve recombinations not of the occupants of actual or possible space–times themselves, but rather of their *duplicates*:

> I cannot altogether accept the formulation: anything can exist with anything. For I think the worlds do not overlap, hence each thing is part of only one of them. A dragon from one world and a unicorn from a second world do not themselves coexist either in the dragon's world, or in the unicorn's world, or in a third world. An attached head does not reappear as a separated head in some other world, because it does not reappear at all in any other world. . . . It is right to formulate our principle of recombination in terms of similarity. . . . But extrinsic similarity is irrelevant here, so . . . I should say that a *duplicate* of the dragon and a *duplicate* of the unicorn coexist at some world, and that the attached talking head has at some world a separated duplicate. (1986, 88–9)

Hume-intrinsicality and duplication form a tight circle here, with (as a first pass) duplicates sharing all Hume-intrinsic properties, and Hume-intrinsic properties being shared by all duplicates. More precisely, according to Lewis, duplicates share their "perfectly natural" properties, where natural properties are supposed to make for greater objective resemblance among their possessing particulars, and where perfectly natural properties are assumed to be "basic intrinsic" – that is, to be fundamental Hume-intrinsic properties; in reverse, fundamental Hume-intrinsic properties are again those shared by all duplicates. Hence notwithstanding Lewis's case studies of dragons and talking heads, the elements guaranteed to be subject to recombination are in the first instance Hume-intrinsic duplicates of points (or regions) instantiating local qualities, or intrinsic duplicates of entities built up out of such points or regions. (The question of how Lewis aims to accommodate possibilities involving macro-level entities will be revisited in section 10.3.)

Fourth, and finally, the principle requires a proviso blocking unlimited recombinations of duplicates from being "too big" to fit in single space–time continuum:

> Our principle . . . requires a proviso: "size and shape permitting". The only limit on the extent to which a world can be filled with duplicates of possible individuals is that the parts of a world must be able to fit together within some possible size and shape of spacetime. Apart from that, anything can coexist with anything, and anything can fail to coexist with anything. (1986, 89–90)

Putting Lewis's primary statement and associated qualifications together, we arrive at the following combinatorial principle:

> *HD (L-combinatorialism)*: Every occupant of spacetime existing at any actual or possible world is such that one of its duplicates (sharing all Hume-intrinsic properties) can coexist with one of the duplicate(s) of any non-overlapping occupant(s) of spacetime existing at any actual or possible world(s), or fail to coexist with one of the duplicate(s) of any occupant(s) of spacetime existing at any actual or possible world(s).

The principle requires that there be worlds enough to accommodate all the combinations; hence it is that it acts as a principled generator of the space of metaphysical modality. Given Lewis's Humean inclinations, he no doubt found HD (L-combinatorialism) intuitively plausible. Most importantly, however, Lewis invokes HD (L-combinatorialism) as sufficing to express or ensure that there are no gaps in metaphysical modal space.

In particular, and importantly for Lewis's larger project (that is, his "sustained campaign" in favor of Humean supervenience), HD's application here confirms and moreover provides needed support to Lewis's supposition that the laws of nature are contingent. We saw earlier that contingentism is plausible, given Lewis's understanding of laws as best systematizations of the regularities, but this result hinges on whether worlds having sufficiently different spatiotemporal distributions of Hume-intrinsic properties are genuinely possible. Indeed, such worlds are possible, by the lights of Lewis's HD (L-combinatorialism). As Lewis notes:

> Another use of my principle is to settle – or as opponents might say, to beg – the question of whether laws of nature are strictly necessary. They are not; or at least laws that constrain what can coexist in different positions are not. (1986, 91)

More generally, the principle effectively generalizes Hume's original application of his dictum to apply to any connections between spatiotemporally (wholly) distinct entities whatsoever:

> It is no surprise that my principle [of recombination] prohibits strictly necessary connections between distinct existences. What I have done is to take a Humean view about laws and causation, and use it instead as a thesis about possibility. Same thesis, different emphasis. (91)

10.2.2 Prima Facie Advantages of Lewis's Combinatorialism

There are three prima facie advantages of Lewis's combinatorialism, understood as involving HD (L-combinatorialism) against the backdrop of concrete modal realism.

First, as above it would be desirable if the space of metaphysical modality were generated in something like the principled and elegant way that the space of syntactically logical modality is generated by the principle of syntactic consistency. On a combinatorial account we have a single principle that has prima facie promise of doing this job.

Second, given that HD is in the business of denying necessary connections (to speak in terms that, as I'll next discuss, appear to be dischargeable), one might be prima facie confident, as Lewis was, that the metaphysical modal space generated by this principle would not leave any gaps.

One might be concerned here that, even granting (what I'll shortly question) that HD (L-combinatorialism) leaves no gaps in metaphysical modal space, nonetheless the principle *overshoots* the desired space. Indeed, by far the most common criticism of Lewis's combinatorialism is that it lets in possibilities that are not in fact such (allowing, e.g., that I might exist at a world though my mother never existed there, contra Kripkean origin essentialism, or that massy entities might repel, contra dispositional essentialism).[9] There is a case to be made, however, that undergeneration is a worse sin than overgeneration in a theory, since contextual, quantificational, and other resources exist for restricting the space in such a way that the letter, if not the spirit, of the objection is accommodated. Hence Lewis refines his view in counterpart-theoretic terms to allow for contexts conforming to the essentialist's constraints, and Schaffer (2005) appeals to Kratzer's (1977) account of quantifiers as multiply ambiguous, in the course of responding to overgeneration concerns. Given such resources, the main thing, one might reasonably think, is to ensure that all the desired genuine possibilities are generated in principled fashion; extras can be excluded in one or other semi-principled fashion. Correspondingly, insofar as HD (L-combinatorialism) denies any and all necessary connections between wholly distinct occupants of space–time, one might think that this principle at worse overshoots in a treatable fashion, hence has promise of being an *extensionally adequate* generator of metaphysical modal space.

The first and second prima facie advantages are shared by other combinatorial accounts, which coincide in citing something like HD, but for which the combinatorial elements are different, as with Armstrong's (1989) account on which the base elements are universals.

The third prima facie advantage appears to be unique to Lewis's account, however, and reflects another way in which Lewis's appeal to HD is deeply in the Humean vein. The potential advantage here lies in HD (combinatorialism) appearing to serve, when implemented inside the framework of Lewis's concrete modal realism, as the basis for a distinctively *reductive* account of modality. Above we observed that contemporary proponents of HD typically do not follow Hume in rejecting IBE as a warranted mode of inference. Where contemporary proponents – in particular, Lewis – typically *do* follow Hume is in their distaste for ontologically irreducible modality, notwithstanding that, for Lewis and others, this scruple is born of metaphysical rather than epistemological considerations. Modality is to be analyzed, or at least accounted for, in non-modal terms.[10] But if so, the question arises: how are we to understand the 'can' that enters into HD (combinatorialism)?

Lewis's concrete modal realism provides him with an answer; namely, that the 'can' is to be analyzed as an 'is' – the 'is' of extension across the available space of concrete possible worlds. To say that any occupants of any space–time can exist, or not exist, with any occupants of any space–time, is to say that duplicates of these occupants *do* exist, or *do not* exist, together in some concrete possible world. As such, possibility is ultimately analyzed in purely extensional terms. Note that forms of combinatorialism, such as Armstrong's, on which the combinatorial elements are universals or some other non-extensional phenomena, do not so clearly discharge the modal implication associated with

their combinatorial element, and indeed, Lewis's primary objection to Armstrong's view was that it was not appropriately reductive, in requiring appeal to a notion of consistency in order to construct "ersatz" possible worlds from universals or other abstracta (see Lewis 1992).

Lewis's HD-based combinatorial approach to modality thus has promise of generating the space of metaphysical modality in a way that is not just principled and extensionally correct, but reductive. It may be that non-combinatorialist approaches can correctly draw the boundaries of metaphysical modal space; here we would have to attend to details. But whatever the details, it's hard to see how a non-combinatorial approach might be as promising as is HD (L-combinatorialism) in either the first or third respects. To start with the question of reduction: pragmatic, conventionalist, dispositionalist, or essentialist approaches will likely advert to modal notions (such that, e.g., what is possible is what we find useful to take to be possible, or is consistent with the natures of the entities involved, where the notion of a nature has immediate, potentially irreducible, modal implications). Not everyone has reductive ambitions, but supposing one does, Lewis's HD-based combinatorialism appears to be the best game in town; and even those not inclined towards modal reductionism can acknowledge that *if* one can generate the space of metaphysical modality in non-modal terms, that would be a win from the perspective of ontological and/or ideological parsimony.

More importantly, in my view, is that existing non-combinatorial accounts do not begin to approach an HD-based combinatorial account in terms of providing a principled and systematic basis for generating metaphysical modal space. Even supposing a single generative principle can be associated with a given such account, actually implementing the account will require – what? Canvassing conventionalist intuitions? Asking the scientists? The end of metaphysical inquiry? Of course, the truth about metaphysical modality might be messy. But if it weren't, that would be incredibly useful, especially since so much philosophy requires that we be able to make informed judgments about what is or is not possible.

10.3 Undergeneration Concerns for Lewis's Combinatorialism

Prima facie, Lewis's combinatorialism promises to provide a *principled*, *extensionally adequate*, and *reductive* generator of metaphysical modal space. But is the promise really fulfilled?

As prefigured, I will argue that this promise is undermined by attention to a variety of ways in which Lewis's combinatorialism undergenerates the space of metaphysical modality, as pertaining, in particular, to broadly scientific entities.

10.3.1 Spatiotemporally Overlapping Hume-Intrinsic Fundamenta

Lewis's combinatorialism applies only to (duplicates of) non-overlapping occupants of space–time ("at least provided they occupy distinct spatiotemporal positions"; 1986, 88). Hence, as it stands, HD (L-combinatorialism) does not specify what is or is not possible for spatiotemporally overlapping entities; the principle is simply silent on the matter.

Is this a significant failure of plenitude? In the next subsection, I'll consider this question as directed at possibilities involving overlapping macro-level objects or properties. Here I'll focus on possibilities involving overlapping Hume-intrinsic fundamenta, for which the answer appears to be "yes," since, after all, possibilities can substantively differ as regards such entities. Some objects may overlap (e.g. bosons), but some may not (e.g. fermions). Some objects and properties may overlap (e.g. electrons and negative charge), but some may not (e.g. electrons and positive charge, being square, or being prime).[11] Most importantly, and most uncontroversially, some properties – negative charge and mass – may overlap (be coinstantiated), others – negative and positive charge – may not (and

more generally, proponents and opponents of HD alike agree that determinates falling under the same determinable can't overlap). Both opponents and proponents of HD typically allow that there may be such substantive modal facts about overlapping entities, and the acceptance of restrictions on overlapping determinates is especially common. Hence, for example, in discussing how HD (L-combinatorialism) supports the contingency of laws of nature, Lewis qualifies: "...perhaps with the exception of laws constraining what can coexist at a single position, for instance the law (if such it be) that nothing is both positive and negative in charge" (1986, 91). And later, when assessing pictorial ersatzism, he says: "[W]e noticed that it is all too easy to *say* that the same particle is both positive and negative in charge. But if in fact these are incompatible determinates...then nothing whatever...has them both" (1986, 168). Interestingly, however, Lewis never discusses, head-on, the fact that his own recombinatorial principle fails to apply to such cases. Of course, in remaining silent on the cases Lewis's view doesn't explicitly fall into error; but it remains that as it stands HD (L-combinatorialism) undergenerates the space of metaphysical modality, and so fails to be extensionally adequate.

One might wonder whether HD might be extended, one way or another, to appropriately treat possibilities for overlapping fundamenta. Doing so requires a notion of "distinct" or "wholly distinct" different from that applying to spatiotemporally non-overlapping entities; it is unclear, however, whether any alternative understanding will do the trick.[12] Mere numerical distinctness results in an extreme version of HD, which in addition to incorrectly deeming it possible that fermions are colocated, that electrons are positively charged, and that some fundamental entity is both positively and negatively charged, would more generally render it possible for sets to have different (individual) members, fusions to have different (individual) parts, and so on. Lewis, it seems, didn't have such a weak notion of distinctness in mind, at least so far as HD (L-combinatorialism) is concerned, since as above he seems willing to allow that there could be barriers to coinstantiation of numerically distinct properties. Taking this route thus doesn't resolve the problem of extensional inadequacy.

Certain other notions of distinctness (besides spatiotemporal non-overlap and numerical distinctness) are given in modal terms, according to which entities are wholly distinct if it is possible for one or both to exist without the other's existing. But a modal characterization of distinctness won't do in a context where HD is being invoked in order to *generate* the space of possibilities. Taking this route to overcoming undergeneration would thus undermine the reductive status of Lewis's combinatorialism.

The remaining and initially promising suggestion takes the relevant notion of distinctness to involve constitution, such that (in particular, spatiotemporally overlapping) entities are wholly distinct just in case neither at all constitutes the other, with HD (L-combinatorialism) extended accordingly. This strategy faces four difficulties, however.

First, a constitution-based extension of HD (L-combinatorialism) only partly overcomes extensional inadequacy. It will correctly deem it impossible that electrons fail to be negatively charged (since electrons are partly constituted by negative charge), and perhaps it will also explain why determinates of a single determinable cannot overlap, if these are each partly constituted by a single determinable. But it will not explain why an electron cannot also be positively charged, prime, or square – after all, in general it is possible for entities to have properties that do not enter into constituting them. Nor will it distinguish between the possibilities for overlap among bosons and fermions, since while bosons do not constitute one other, neither do fermions. Second, invocations of constitution as a basis for necessary connections may tacitly reintroduce necessary connections between wholly distinct entities, contra HD (L-combinatorialism). As I have argued (Wilson 2010b), in many cases, the best explanation of why we are justified in taking there to be necessary constitutional connections – again, accepted by Humeans and non-Humeans alike – presupposes that there are necessary causal connections, contra the core applications of HD, and more specifically contra Lewis's

intended use of HD (L-combinatorialism) (not to mention his assumption of Humean supervenience as involving Hume-intrinsic fundamenta) as establishing the contingency of laws of nature. So, for example, the best explanation of why we are justified in accepting that necessarily, electrons are negatively charged – accepted by both friends and foes of HD – adverts to there being a modally stable overlap in the causal profiles of the two entities, contrary to HD (L-combinatorialism). Third, a constitution-based understanding of whole distinctness, and associated version of HD (L-combinatorialism), relies on our having some principled means of saying when some entities do or do not constitute some others, but we do not have any such principled means: such investigations are both methodologically opaque and tangled up with various of one's other commitments. If applications of HD (L-combinatorialism) to cases of spatiotemporally overlapping fundamenta must rely on the outcomes of investigations into the constitution of such entities, then its claim to provide a principled and elegant generator of the space of metaphysical modality is undermined. Fourth, a constitution-based understanding also threatens to undermine the claim that HD (L-combinatorialism) provides the basis for a reductive account of modality, for investigations into questions of constitution typically proceed by consideration of what is possible or necessary for a given entity.

Summing up: the failure of Lewis's combinatorialism to specify what is possible or necessary for spatiotemporally overlapping Hume-intrinsic fundamenta poses a serious problem for the claim that HD (L-combinatorialism) is an extensionally adequate generator of metaphysical modal space. Moreover, attempts to overcome this deficiency by extending the combinatorial principle in one or other fashion either fail to overcome extensional inadequacy, undermine the claim that HD (L-combinatorialism) is a principled generator of metaphysical modal space, undermine the reductive aspirations of Lewis's account, or indirectly entail the falsity of HD. A similar menu of concerns will attach to the other cases of undergeneration that I'll now mention.

10.3.2 Undergenerated Macro-Possibilities

Let us turn now to possibilities concerning macro-entities. It is possible, I hope you'll agree, that there be a plaid kangaroo. Is this possibility generated by HD (L-combinatorialism)? Not without further assumptions. HD (L-combinatorialism) generates "juxtapositional" possibilities – for example, it generates the possibility that there be a kangaroo wearing (that is, appropriately proximate to) a plaid coat. But the possibility of a plaid kangaroo is not a juxtapositional possibility, but rather requires that the property of being plaid somehow overlap (that is, be instantiated in) a kangaroo.

What further assumptions are needed? Here Lewis will appeal to his doctrine of Humean supervenience, according to which "all else" supervenes on the distribution of Hume-intrinsic qualities. Recall: "[W]e have an arrangement of qualities. And that is all. There is no difference without difference in the arrangement of qualities. All else supervenes on that" (1986, ix). As such, and notwithstanding Lewis's case studies involving kangaroos, dragons, talking heads and other macro-entities, the possibilities to which HD (L-combinatorialism) is in the first instance directed are possibilities involving juxtapositions of elements in (actual or possible) Humean supervenience bases. Hence when the proponent of HD (L-combinatorialism) speaks of what is possible for a kangaroo, they are *really* speaking of what is possible for the presumed supervenience base of kangaroos; and in saying that it is possible that there be a plaid kangaroo, they are saying, of entities in that presumed supervenience base, that (duplicates of) these entities might be configured so as to provide a supervenience base for a plaid kangaroo.

Before assessing this strategy, a qualification is in order, reflecting that the supposition that there is such a base doesn't follow from HD (L-combinatorialism) alone. This principle guarantees that (duplicates of) the entities (actual or possible) at each world may exist together, or exist apart, at a

world; but it doesn't guarantee that coexisting duplicates will stand in the relations requisite unto providing a supervenience base for, for example, a plaid kangaroo. Hence Divers and Melia note:

> [Lewis's stated] principle of recombination is insufficient to generate worlds where donkeys talk or where there are blue swans. Grant that there is a set of actual particles such that were those particles arranged in a certain way they would constitute a talking donkey. It is not enough for a world to represent that possibility that it should contain a duplicate of each such particle, for obviously such particles could exist, scattered to the corners of the universe, without constituting a talking donkey. (2002, 16)

To overcome this sort of undergeneration, they rather suggest formulating the recombinatorial principle in a way that requires that there is a distinct possibility for *every* way of spatiotemporally relating the relevant duplicates:

> For any individuals x_1, x_2, . . . , x_n there is a world containing any number of duplicates of each, if there is a spacetime big enough to hold them all, and such that for any spatiotemporal relation the duplicates in question stand in that relation.[13]

By doing so, they claim that "we achieve the effect of capturing those arrangements of the particles in question that constitute a talking donkey" (16).

The need to ensure that HD (L-combinatorialism) generates duplicates that are spatiotemporally related in the right way seems correct (for simplicity I won't carry this qualification through), though as I'll argue down the line there is a difficulty for Lewis's reductive aspirations here, as regards what metaphysically determines that some Hume-intrinsic goings-on serve as the basis for a macro-entity of a given type. In any case, more important than the need to quantify over spatiotemporal relations in the combinatorial principle is the tacit assumption that Humean supervenience must be presupposed if HD (L-combinatorialism) is to have any chance of capturing the sort of macro-level possibilities in question.

This assumption undermines the claim that HD (L-combinatorialism) provides a principled generator of the space of metaphysical modality. It is not HD (L-combinatorialism) that generates the space, but this principle in tandem with another, substantive principle, according to which "all else supervenes" on the distribution of Hume-intrinsic qualities at space–time points. When Lewis looks to HD (L-combinatorialism) he is looking not just for a principle that will generate all the possible worlds, but for a principle that will generate all the *possibilities*. What HD (L-combinatorialism) generates is worlds containing recombinations of Hume-intrinsic fundamenta. Do any such worlds contain plaid kangaroos? Well, that depends on whether Humean supervenience is true. After all (to focus on one of their salient features), kangaroos are sentient creatures. If consciousness is a robustly emergent property of complex combinations of micro-level (here, Hume-intrinsic) goings-on, then HD (L-combinatorialism) alone would not axiomatically generate a world of the desired plaid, kangaroo-y variety.

So the assumption of Humean supervenience is in fact an additional principle, that must be added to HD (L-combinatorialism) if Lewis's combinatorialism is to be extensionally adequate. That two principles, rather than one, are required to generate metaphysical modal space to some extent undermines the claim that HD (L-combinatorialism) is a principled generator of this space. But the failure of systematicity here goes far beyond multiplicity. One can talk oneself into thinking of combinatorial principles as broadly mathematical means of expressing (or aiming to express) an extremum along the spectrum of conceptions of metaphysical modality – namely, a conception of such modality as highly unrestricted (exceeded only by broadly syntactic logical modality), such that everything not forbidden is possible. But Humean supervenience is another sort of principle altogether. Its truth is

not appropriately seen as a broadly axiomatic stipulation; rather, it is an empirical thesis, contingent on whether or not any properties of macro-entities are constituted or otherwise grounded in fundamental configurational interactions (see McLaughlin 1992). As is indicated by the ongoing disputes in the physicalism debates over the status of mentality and other features whose physical acceptability remains controversial, the confirmation (or disconfirmation) of this thesis is a messy, unprincipled, broadly piecemeal affair. Moreover, given the broader goal of overcoming undergeneration, it won't do to simply refer to a "representative sample," since any exception will break the rule. Undergeneration is thus avoided only by undermining the claim that HD (L-combinatorialism) is a principled (systematic, elegant) generator of metaphysical modal space.[14]

The considerations here also indirectly undermine the reductive aspirations of Lewis's account. To start, as above, Lewis's combinatorialism appears to have a reductive advantage over alternative forms of combinatorialism (e.g. Armstrong's) when it comes to the construction of possible worlds, representing the possibilities for the combinatorial fundamenta (sometimes said to be "true of" a given world). Some also claim that Lewis's account is the only one able to reductively accommodate possibilities pertaining to macro-entities (sometimes said to be "true in" a world). So, for example, here Sider argues that a sort of "best-case" alternative combinatorial account will be modally circular:

> Identifying worlds with sets of space-time points may eliminate modality from the definition of "possible world," but as Lewis has argued, modality reappears in the definition of "true in". What would it mean to say that it is true in a certain set, S, of space-time points that there exists a talking donkey? . . . If we could analyze "talking donkey" in terms of occupied points of space-time then we could determine precisely which patterns of occupation would suffice for the existence of a talking donkey, and then we could say that it is true in S that there is a talking donkey iff S contains one of these patterns. But no one knows how to provide this sort of analysis of "talking donkey". Moreover, a general analysis of modality requires a general definition of "proposition p is true in set S" for arbitrary propositions p; a series of one-off definitions for a few chosen propositions is no progress toward a general analysis. We might define "p is true in possible world w" as meaning "necessarily: if all and only the points in w are occupied by matter then p is true." But this definition uses necessity. No other definition seems available; "true in", therefore, renders the account of modality circular. (2003, 189)

By way of contrast, Sider claims, a concrete modal realist like Lewis does not face a circularity concern. Why not? Effectively, the strategy here aims to apply Lewis's "extensional gambit", according to which what "can" or "must" be the case is to be analyzed in terms of what "is" the case in some or all concrete possible worlds. In particular, on the assumption that Humean supervenience is true, whether a given macro-entity (e.g. a kangaroo) exists at a world and has a given property (e.g. being plaid) will just be true, or not – if the appropriate base exists, the associated macro goings-on will exist, as an extensional matter of fact. Moreover, the seemingly modal implications of the supervenience conditional will also be extensionally discharged, now across worlds rather than within a world: so long as every world containing (a duplicate of) such a base also contains (an overlapping duplicate of) the associated macro-entity, Lewis is, it seems, reductively good to go.

But, as I'll now argue, Lewis's treatment of possibilities involving macro-entities ultimately does not avoid commitment to irreducible modality. There is an interesting contrast in this respect with the case of Hume-intrinsic fundamenta, so let me start by discussing this case. Suppose that it is possible that some actual electron be one foot left of where it actually is, and consider a non-actual concrete world that is supposed to provide a ground for this claim. What ensures, metaphysically speaking, that some Hume-intrinsic entity at this non-actual world is appropriately taken to be a representative, so to speak, of the electron, such that the situation of the duplicate

appropriately bears on what is possible for the actual electron? Here the answer is easy: this is ensured by the duplicate's being *exactly intrinsically similar* to the actual electron. More generally, the relevance of some non-actual Hume-intrinsic fundamenta to what is possible or necessary for some actual Hume-intrinsic fundamenta is ensured by exact intrinsic similarity of the fundamenta in question.[15]

Not so, at least typically, for possibilities involving macro-entities. In the actual world, there are brown kangaroos, grounded, on Lewis's operative assumption, in some Hume-intrinsic supervenience base. Given this, which non-actual worlds are relevant to the possibility that there be a plaid kangaroo? These cannot be worlds containing exact intrinsic duplicates of (the appropriately related Humean supervenience base for) actual kangaroos, since these wouldn't be *plaid* kangaroos: kangaroos of different colors must be at least somewhat intrinsically different.[16] So the non-actual worlds grounding the macro-possibility in question will be ones that are somewhat, but not exactly, intrinsically similar to actual brown kangaroos. But which respects of, and how much, similarity is required? More generally: which non-actual goings-on give rise to kangaroos, and which don't?

Sider (and also Divers and Melia) seems to assume that an extensionalist answer to this question can be given, to the effect that we can just read off of the space of concrete possible worlds which Hume-intrinsic goings-on serve as supervenience bases for non-actual kangaroos. But that's incorrect. Even if its true that *if* some Hume-intrinsic goings-on serve as a kangaroo base, *then* from the existence of the base it will "extensionally" follow (without any need, in particular, for an analysis of kangaroos in Hume-intrinsic terms) that a kangaroo exists, the question here at issue remains: *which* Hume-intrinsic goings-on serve as bases for (e.g. plaid) kangaroos differing from actual kangaroos? One might take this to be primitive, but this answer would leave open that some irreducible modality was involved. Moreover, we are now in position to give a tu quoque against Sider and his associated defense of the reductive status of Lewis's combinatorialism. For even if one could give a non-primitive answer to the question of whether some Hume-intrinsic goings-on serve as a basis for a non-actual kangaroo, "what is needed for a general account of modality" is a *general* account of what it takes for some Hume-intrinsic goings-on to serve as a basis for a non-actual macro-entity of a given type. And here it seems that the only available account is one according to which the relevant Hume-intrinsic goings-on are those serving as the basis for a *possible* macro-entity of the type, where the notion of possibility here is perforce intensional rather than extensional (as per the first point). Hence proper accommodation of possibilities pertaining to macro-entities renders Lewis's combinatorialism circular.

Here again, then, HD (L-combinatorialism) fails to be a complete generator of the space of metaphysical modality, and available ways of closing the gap undermine the supposed methodological advantage and/or the reductive credentials of this principle.

10.3.3 The Presupposition of Spatiotemporal Fundamentality

My third and final undergeneration concern stems from Lewis's supposition that the elements whose recombination is supposed to serve as a generative basis for the space of metaphysical modality are occupants of actual or possible space–times. This assumption is problematic, since it is unclear that the most fundamental elements are occupants of space–time, as opposed to some more abstract space – most saliently, configuration space (see Paul 2012; Ney 2012, and others). If this last is correct, then a comparatively vast array of possibilities – all those associated with spaces in some or other respects different from ordinary spatiotemporal space) – are simply left out of Lewis's account. Lewis might try to maintain that these possibilities are not genuine, but in doing so he would be on shaky ground, by his own lights, for he counsels us to look to the sciences to get a handle on what is

fundamental (or "perfectly natural"), and it is the sciences that have raised to salience the possibility of non-spatiotemporal spaces.

One might try, as with the previous undergeneration of possibilities associated with overlapping entities, to expand HD (L-combinatorialism) so as to apply to occupants, "overlapping" or not, of more abstract spaces. Here an approach along lines of Armstrong's universal-based version of combinatorialism seems potentially promising, though there remain questions about how to individuate universals (or more to the point, their instantiations) as wholly distinct or not. But Lewis cannot accept this strategy for overcoming undergeneration, since as previously noted he rejects Armstrong's combinatorialist account as failing to appropriately accommodate reductionism.

Indeed, Lewis's reductive aim is threatened, no matter how HD (L-combinatorialism) might be expanded to accommodate configuration and other more abstract spaces. As above, while Lewis follows Hume in thinking that there is no irreducible modality, a mere appeal to HD as a generative principle does not in itself ensure that modality is so reducible, since HD itself involves modal vocabulary. It was specifically Lewis's appeal to HD as applying to *occupants of space–time*, in the context of his *concrete modal realism*, that ensured that this modal vocabulary was discharged, as really involving merely extensional facts. Once the combinatorial principle is applied to more abstract elements, however, which are not (at least in the first instance) occupants of space–time but rather of some more general space, it is no longer clear that Lewis's extensional gambit applies.

10.4 The End Game

I have argued that HD is not motivated as an inference to the best explanation of the range of metaphysical modality – that is, as presupposed by Lewis's combinatorialism. For, I have argued, Lewis's account undergenerates the space in several important respects, and available means of narrowing the gaps in metaphysical modal space either (1) fail to close the gap, leaving HD (L-combinatorialism) extensionally inadequate; (2) require supplementing HD (L-combinatorialism) with principles, such as a constitution-based understanding of 'wholly distinct' or the thesis of Humean supervenience, which undermine HD (L-combinatorialism)'s claim to be a *principled* generator of metaphysical modal space; (3) undermine the reductive status of HD (L-combinatorialism); or (4) undermine the truth of HD, as applied, in particular, to the case of causal or nomological connections.

This result does not entirely rule out that HD might be indirectly motivated as a combinatorial generator of the range of metaphysical modal space. Two considerations are especially salient as regards the "end game," so to speak.

First, there exist other combinatorial accounts which presuppose a version of HD – most notably, Armstrong's combinatorial account, on which the combinatorial elements are the fundamental universals and particulars entering into atomic states of affairs, and possible worlds are understood as maximally consistent states of affairs. Though I cannot enter into details here, such a combinatorial account has some promise of systematically generating at least some of the possibilities left ungenerated on Lewis's account.

As above, Lewis rejects Armstrong's account as not genuinely modally reductive, in appealing to a notion of broadly semantic (as opposed to merely logically syntactic) consistency. One might worry that, as such, Armstrong's combinatorialism does not really motivate HD, understood as incorporating Hume's rejection of irreducible modality. This worry can be addressed, however. Hume's rejection of irreducible modality targeted, in the first instance, powerful essences or other purportedly suspicious or inaccessible phenomena involving broadly scientific entities. He was perfectly happy with

semantic consistency as a generator of possibilities. Recall Hume's initial statement of his version of HD in the *Treatise*:

> There is no object, which implies the existence of any other if we consider these objects in themselves, and never look beyond the ideas which we form of them. Such an inference wou'd amount to knowledge, and wou'd imply the absolute contradiction and impossibility of conceiving any thing different. (1978 [1739], Part III, S VI)

Hume doesn't think, in the causal cases he is considering, that there is any such "contradiction" in the relevant states of affairs, but the present point is simply that he doesn't seem opposed to semantic consistency (of the sort presumably operative in instances of conceiving) playing a role in generating the relevant space of possibility. So while one may aim to out-Hume Hume, as Lewis does, in characterizing metaphysical modal space in completely non-modal terms, doing so isn't a prerequisite of appropriately implementing HD as a combinatorial principle, or so it seems to me. So for all that I have argued here, HD might be motivated as presupposed by a combinatorial account of modality other than Lewis's, even if that account fails to be strongly modally reductive.

Beyond combinatorialism, there are many other accounts of metaphysical modality that need to be considered, not just with an eye to extensional adequacy but also to the clear desideratum to provide an account of the space of metaphysical modality that is *principled*, in something like the way the space of logically syntactic possibility is generated by the single principle of syntactic consistency. And here one might think that the prospects of doing without combinatorialism are not so promising.

So, to consider just one example, take Kit Fine's (2002) conjecture that modal facts can be analyzed as follows: for it to be necessary that p just is for there to be some things, X, such that p holds in virtue of the natures of the Xs. Here the suggestion might be that the space of metaphysical modality is generated by a principle to the effect that any scenario compatible with the natures of the entities in the scenario is possible. Such a "nature" or "essence"-based view appeals to broadly semantic consistency; but so does the remaining combinatorialist contender, so this appeal doesn't count as a serious disadvantage. A more telling disadvantage is that a nature- or essence-based view is only superficially principled. Essences are often thought to be mysterian; but the concern at issue here is more specifically that – at least insofar as the essence of essence remains elusive – this generator of metaphysical modal space really adverts to a huge multiplicity of identity and individuation conditions, whose methodology and metaphysical standing remain opaque, both as directed at particular cases, and in general. As with the appeal to constitution as a means of extending HD (L-combinatorialism) to accommodate possibilities for spatially overlapping entities, the appeal to essence seems to count more as an admission of defeat in the goal of providing a principled basis for metaphysical modal space, than it counts as a way of satisfying this goal. Supposing so, that might throw us back towards some form of combinatorialism, and hence to some form of HD.

Second, at the end of the day there is one other consideration that might lead us to endorse HD, as presupposed by Lewis's account of combinatorialism. While, as I've here tried to show, Lewis's account of combinatorialism has its problems, other accounts also have their problems. And even if *on this specific issue* Lewis's account, and associated appeal to HD, does not come out ahead, it might be that *all things considered* Lewis's overall framework motivates HD, as presupposed not just in his account of the space of metaphysical modality but moreover, as we have seen, as presupposed in nearly every important aspect of his systematic philosophy. And so far as the end game of deciding which systematic philosophy is best,[17] Lewis's views are very much still in the running.

Acknowledgments

Thanks to Christopher Gibilisco, Benj Hellie, Alastair Wilson, participants at the Fifth Annual Chambers Philosophy Conference ('Necessity') at the University of Nebraska, and students in my University of Toronto seminars on Hume's dictum for helpful comments and discussion.

Notes

1 See, for example, Armstrong 1983, Lewis 1992, Paull and Sider 1992, Bennett 2004, and Moyer 2008.
2 See Wilson 2010a, 2010b, and 2014.
3 See Wilson 2010a for detailed discussion of these interpretive options (including, e.g., as applying to individuals or types – a distinction that nothing turns on here) for contemporary versions of HD.
4 In cases of ties, all the closest P-worlds must be Q-worlds; in cases where there is (are) no closest P-world(s), the usual strategy is to require that all P-worlds within a certain range of proximity be Q-worlds.
5 In brief, the strategy of Wilson (2014) is as follows. Lewis's (and Schaffer's) main cited motivations for endorsing a "miracle-based" account of comparative overall similarity are, first, that some salient contexts of counterfactual evaluation presuppose *CF asymmetry*, and, second, that accounts of counterfactuals failing to presuppose or impose *CF asymmetry* are epistemologically problematic, in that under conditions of determinism, the variations in initial micro-conditions needed to implement a given counterfactual antecedent would result in so many changes to ensuing macro-states that evaluation of CFs would be rendered practically impossible. Against the first reason, I argue that no non-artificial contexts involving counterfactual reasoning presuppose *CF asymmetry*; against the second, I observe that micro-variation associated with variations in initial conditions is compatible, in principle, with significant similarity as regards macroscopic states of affairs – enough, in particular, to allow counterfactuals to be appropriately evaluated.
6 I say "instance" since there are other HD-presupposing implementations of combinatorial accounts (e.g. Armstrong's) on which the base elements, and corresponding understanding of "whole distinctness," are different from Lewis's. Assessment of alternative combinatorial accounts must await another day.
7 The reference to HDs being taken to be a "broadly axiomatic" generator of metaphysical modal space (I won't always carry this qualifier through) is meant to flag that the principle is intended to characterize certain worlds (namely, the metaphysically possible ones) in something like the way that the Peano axioms are intended to characterize certain mathematical entities (namely, the integers), as opposed to metaphysically cause or otherwise bring into existence the worlds in the space, or to select these worlds from a pre-existing broader space. Lewis clearly also sees HD as giving us an epistemic handle on which worlds exist; while in this paper the primary focus is on HD's usefulness as a metaphysical principle, the question of HD's epistemic usefulness will come up in section 10.3.2.
8 A related concern, not discussed by Lewis, is that the two-part principle does little more than encode, for worlds or parts of worlds, the operative assumption that modal claims involve quantification over possible worlds (such that, e.g., Ps being possible is understood in terms of Ps being true in some possible world). The two-part principle thus provides a translation strategy for modal claims, but does not provide any independent handle on which worlds are in the space.
9 See, e.g., Carroll 1990, Shoemaker 1998, Bird 2007, and many others.
10 That said, there are stronger and weaker ways of interpreting (hence accommodating) Hume's distaste for irreducible modality, as I'll discuss in section 10.4.
11 Here I assume that overlap is a necessary condition for property instantiation; if not, then it remains that such modal differences are not generated by HD (L-combinatorialism).
12 See Wilson (2010a) for detailed discussion of different notions of distinctness that might enter into formulating HD, and assessment of the resulting theses.

13 This is not perspicuously phrased (as it stands its interpretation is compatible with the claim that the duplicates in question exist at a single world and somehow stand in every single spatiotemporal relation), but the idea is clear enough.

14 Indeed, as Christopher Gibilisco pointed out, undergeneration will not be avoided, even at this price. For, as Lewis qualifies, "I concede that Humean supervenience is at best a contingent truth. Two worlds might indeed differ only in unHumean ways, if one or both of them is a world where Humean supervenience fails" (Lewis, 1987, x); and again: "Humean Supervenience is meant to be contingent: it say that among *worlds like ours*, no two differ without difference in the arrangement of qualities" (Lewis, 1994, 474; emphasis in text). But if the recombination principle requires Humean supervenience in order to fill some gaps in logical space, then Lewis's combinatorialism will rule out the non-Humean worlds that he thinks are possible, and so introduce other such gaps.

15 Hence it is that, according to Lewis, the perfectly natural properties – that is, perfectly (Hume-intrinsic) fundamental properties – make for perfect objective intrinsic similarity among their possessing particulars (see Lewis 1984, 227, and 1986, 60).

16 Here we are ignoring potential overgeneration concerns stemming from the fact that Lewis's view does not explicitly incorporate resources for the (commonly supposed) historic and relational individuation of species types and tokens.

17 This is assuming it makes sense to ask what systematic philosophy is best, of course. That this makes sense might be questioned for, for example, Carnapian reasons, or on grounds that philosophy is in the business of creatively identifying and rigorously exploring the space of alternative theories (as Benj Hellie suggested), as opposed to honing in on a supposed one true theory.

References

Armstrong, D. (1983). *What is a law of nature?* Cambridge: Cambridge University Press.

Armstrong, D. (1989). *A combinatorial theory of modality*. Cambridge: Cambridge University Press.

Bennett, K. (2004). Global supervenience and dependence. *Philosophy and Phenomenological Research*, 68 (3), 501–29.

Bird, A. (2007). *Nature's metaphysics: laws and properties*. Oxford: Oxford University Press.

Carroll, J. (1990). The Humean tradition. *Philosophical Review*, 99, 185–219.

Divers, J. & Melia, J. (2002). The analytic limit of genuine modal realism. *Mind*, 111, 15–36.

Fine, K. (2002). The varieties of necessity. In T.S. Gendler & J. Hawthorne (Eds.), *Conceivability and possibility*, pp. 253–81. Oxford: Oxford University Press.

Hume, D. (1978). *A treatise of human nature*. Ed. L.A. Selby-Bigge & P.H. Nidditch. Oxford: Clarendon Press. Orig. work 1739–40.

Kratzer, A. (1977). What 'must' and 'can' must and can mean. *Linguistics and Philosophy*, 1, 337–55.

Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, 13, 455–76.

Lewis, D. (1986). *On the Plurality of Worlds*. Oxford: Blackwell.

Lewis, D. (1987). *Collected Papers*, Vol. II. Oxford: Oxford University Press.

Lewis, D. (1992). Armstrong on combinatorial possibility. *Australasian Journal of Philosophy*, 70, 211–24.

Lewis, D. (1994). Humean supervenience debugged. *Mind*, 103, 473–90.

McLaughlin, B. (1992). The rise and fall of British emergentism. In A. Beckerman, H. Flohr & J. Kim (Eds.), *Emergence or reduction? Essays on the prospects of non-reductive physicalism*, pp. 49–93. Berlin: De Gruyter.

Moyer, M. (2008). Weak and global supervenience are strong. *Philosophical Studies*, 138 (1), 125–50.

Ney, A. (2012). The status of our ordinary three dimensions in a quantum universe. *Noûs*, 46, 525–60.

Paul, L.A. (2012). Building the world from its fundamental constituents. *Philosophical Studies*, 158, 221–56.

Paull, R.C. & Sider, T. (1992). In defense of global supervenience. *Philosophy and Phenomenological Research*, 52 (4), 833–53.

Schaffer, J. (2005). Quiddistic knowledge. *Philosophical Studies*, 123, 1–32.
Schaffer, J. (2009). On what grounds what. In D. Chalmers, D. Manley & R. Wasserman (Eds.), *Metametaphysics*, pp. 347–83. Oxford: Oxford University Press.
Shoemaker, S. (1998). Causal and metaphysical necessity. *Pacific Philosophical Quarterly*, 79, 59–77.
Sider, T. (2003). Reductive theories of modality. In M.J. Loux & D.W. Zimmerman (Eds.), *The Oxford Handbook of* Metaphysics, pp. 180–208. Oxford: Oxford University Press.
Stalnaker, R. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory*, pp. 98–112. Oxford: Blackwell.
Swoyer, C. (1982). The nature of natural laws. *Australasian Journal of Philosophy*, 60, 203–23.
Wilson, J. (2010a). What is Hume's dictum, and why believe it? *Philosophy and Phenomenological Research*, 80, 595–637.
Wilson, J. (2010b). From constitutional necessities to causal necessities. In H. Beebee & N. Sabbarton-Leary (Eds.), *Classifying nature: the semantics and metaphysics of natural* kinds, pp. 192–211. New York: Routledge.
Wilson, J. (2014). Hume's dictum and the asymmetry of counterfactual dependence. In A. Wilson (Ed.), *Chance and temporal asymmetry*, pp. 259–79. Oxford: Oxford University Press.

11

Truthmaking

With and Without Counterpart Theory

PHILLIP BRICKER

According to the Truthmaker Principle: every truth has a truthmaker. Spelled out a bit more exactly: for every true proposition, there is some entity whose existence entails, or necessitates, the truth of the proposition. The demand for truthmakers has been championed most vigorously by David Armstrong, who, following C.B. Martin, touts it as a way to keep philosophers honest by requiring that they pay the full ontological cost of their theories.[1] Numerous philosophical views, including versions of phenomenalism, behaviorism, and presentism, would seem to run afoul of the Truthmaker Principle. Moreover, according to Armstrong, the demand for truthmakers provides an alternative to the orthodox Quinean approach to ontological commitment, which, due to its narrow focus on the quantifiers, tends to underestimate the ontological cost of theories.[2] As David Lewis writes: "We can scarcely exaggerate the importance of the demand for truthmakers throughout Armstrong's writings" (Lewis 1992, 215).

Attempts to come to grips with the Truthmaker Principle played a prominent role in Lewis's metaphysical writings over the last 15 years of his career. He first grappled with the principle in "A Comment on Forrest and Armstrong."[3] Then he launched a critique of the principle in reviews of Armstrong's books, *A Combinatorial Theory of Possibility* and *A World of States of Affairs*.[4] Although Lewis agreed that the truth of propositions must somehow be ontologically grounded, the Truthmaker Principle was too strong; it conflicted with two of Lewis's most fundamental metaphysical assumptions, the uniqueness of composition and the Humean denial of necessary connections. Lewis endorsed instead a weaker principle according to which truth supervenes on being: any true proposition is made true by the pattern of instantiation of fundamental, or perfectly natural, properties and relations among particular things.[5] Then, in a series of three papers, among his final writings in metaphysics, he further developed and crystallized his view. In Lewis (2001b) he argued that the Truthmaker Principle, even if accepted, does not support a correspondence theory of truth. In Lewis (2001a), he sharpened his critique of the Truthmaker Principle by showing that it is equivalent to an implausible principle as to how possible worlds must differ. Finally, in Lewis (2003) and Lewis and Rosen (2003), he changed course, noting that his critique of the Truthmaker Principle rested on essentialist assumptions that he, as a counterpart theorist, does not accept. Once freed

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

from those assumptions, a counterpart theorist can accept the Truthmaker Principle after all without buying into unmereological composition and mysterious necessary connections. Ironically, it is Lewis, not Armstrong, who can accept the Truthmaker Principle without paying a hefty metaphysical cost.

The divide over the Truthmaker Principle corresponds to a divide over fundamental ontology. For Armstrong, the demand for truthmakers leads to an ontology of *facts*, or *states of affairs*, thus supporting Wittgenstein's famous Tractarian saying: "the world is the totality of facts, not of things." (I will stick with Armstrong's preferred term 'state of affairs' in what follows.) States of affairs are (somehow) composed of particulars and immanent universals, and are themselves immanent entities that go towards making up the world. How abundant are the states of affairs? I will suppose at a minimum that all states-of-affairs theorists accept a full slate of *atomic* states of affairs: for any atomic proposition $Ra_1a_2\ldots$, there is an atomic state of affairs S such that, necessarily, S exists (obtains) if and only if $Ra_1a_2\ldots$ is true.[6] For Lewis, states of affairs are trouble. Lewis is a *thing theorist*: the world is a *thing*, the *biggest* thing, and every spatiotemporal part of the world, no matter its shape or size, is also a thing. Perhaps in addition to things there are non-spatiotemporal components of things – immanent universals or tropes. Lewis remained agnostic between a nominalist and a realist version of thing theory. But a thing theorist does not admit states of affairs. Both Lewis and Armstrong hold that truth is ontologically grounded. But whereas Armstrong thinks that states of affairs are needed to ontologically ground truth, Lewis holds that an ontology of things (perhaps together with immanent properties and relations) provides ground enough.

Here, in brief outline, is the plan of the following essay. In section 11.1, I introduce the idea of truthmaking, and consider how much truthmaking can be done by *things* without making controversial assumptions in modal metaphysics. In section 11.2, I present Armstrong's account of truthmaking and the states of affairs he thinks are needed to play the role of truthmakers. In section 11.3, I present Lewis's view that the Truthmaker Principle does not make any claim about the notion of truth, and *a fortiori* should not be taken to be a version of the correspondence theory of truth. In section 11.4, I present Lewis's critique of the Truthmaker Principle, and the ontology of states of affairs that (typically) goes with it. In section 11.5, I present Lewis's claim that, as a counterpart theorist, he can accept the Truthmaker Principle after all. Here, however, I think Lewis's position needs to be amended. Finally, in section 11.6, I present and discuss Lewis's alternative to the Truthmaker Principle, the idea that "truth supervenes on being."

11.1 The Theory of Truthmaking

Before considering Lewis's critique of the Truthmaker Principle, we need to bring the principle into sharper focus, and view the principle through Lewis's eyes. Lewis's discussion of the principle (prior to Lewis 2003) was meant to be neutral with respect to controversial metaphysical assumptions, such as his own modal realism, or his counterpart theory. He supposed only that talk of possible worlds made sense, whether worlds were taken to be concrete, or abstract, or fictional; and that entities could be said to exist in multiple possible worlds, whether existing-in-a-world was to be understood in terms of "transworld identity" or counterpart relations. Moreover, as should go without saying, Lewis interpreted the principle in terms that he found intelligible, so without recourse to a primitive grounding, or dependence, or in-virtue-of relation.

First, I ask: how does Lewis understand the truthmaking relation? For Lewis, truthmaking is a modal relation, and is to be understood in terms of strict implication. When an entity makes a proposition true, it is metaphysically impossible for that entity to exist without the proposition being true. Thus, as an initial formulation of the Truthmaker Principle, we have:

(TM) For every true proposition *P*, there exists some entity *T* such that, necessarily, if *T* exists, then *P* is true.

Three things should be noted at the start. First, on this account, the truthmaking relation – a relation between *entities* and propositions – is analyzed in terms of strict implication – a relation between *propositions* and propositions.[7] The analysis requires the assumption that, for any entity, there is a proposition asserting that that entity exists. This assumption, however, is not problematic on an abundant, non-linguistic conception of propositions of the sort that Lewis is supposing (see below). The analysis has the consequence that it is a trivial analytic truth that any (existent) entity is a truthmaker for the proposition that that entity exists. But that seems exactly right.

Second, on this account, the Truthmaker Principle is an assertion of modality *de re*: it involves quantifying into a modal context. It thus presupposes that modality *de re* is coherent (*pace* Quine). In particular, it presupposes that it makes sense to say that an entity exists in multiple worlds. For Lewis, since possible worlds are concrete and do not overlap, modality *de re* must be interpreted using counterpart theory (see section 11.5). But Lewis's discussion of (TM) (prior to Lewis 2003) does not presuppose any particular account of possible worlds or of *de re* representation.

Third, (TM) does not by itself assert that all truths are *ontologically grounded* because the truthmaking relation, as Lewis understands it, does not require truthmakers to be in any way fundamental. Consider the following example. On standard assumptions, I cannot exist without my singleton, nor can my singleton exist without me. Thus, according to (TM), not only am I a truthmaker for the proposition that my singleton exists, but my singleton is a truthmaker for the proposition that I exist; for my singleton is such that, necessarily, if it exists, then I exist. But, surely, the truth that I exist is not *grounded* in the existence of my singleton. If a truthmaker theorist wants to capture the idea that truth is ontologically grounded, she has a choice. She can enhance the notion of a truthmaker, requiring that a proposition *P* is made true by an entity *T* only if both *P* is strictly implied by the existence of *T* and *P* holds *in virtue of T* (cf. Armstrong 2004, 17). (Depending upon how the in-virtue-of relation is understood, the first conjunct may be redundant.) Or, she can keep the truthmaking relation as it is, realizing that something more will have to be added to (TM) to capture the idea that all truths are ontologically grounded. I will follow Lewis in taking the latter course. Lewis's critique of (TM) is not affected by this choice. Arguably, the intuitive notion of *making* a proposition true includes the notion of grounding, but the terminological decision not to so include it won't lead to trouble if it's made clear at the start.

Next I ask: in taking the bearers of truth to be propositions, what conception of proposition does Lewis intend to invoke? For Lewis, there are many equally legitimate conceptions of proposition; which conception is appropriate to a given task depends upon the role that the propositions are required to play, be it semantic, or epistemic, or metaphysical (see Lewis 1986a, 53–5). In the context of seeking truthmakers, what matters is only the *content*, irrespective of how that content is represented, or whether that content is in any way fundamental. When content alone matters, Lewis identifies propositions with classes of possible worlds. But, for his critique of the Truthmaker Principle, Lewis does not want to impose his own view, and assumes only the following. First, propositions are not linguistic entities, nor in any way tied to language; the metaphysical demand for truthmakers should not be limited by the expressive capabilities of actual languages, or even humanly possible languages. Second, propositions are abundant to at least this extent: for any possible world, there is a proposition asserting that that world is actualized. Beyond this, Lewis's assumptions on propositions are standard: the propositions form a Boolean algebra, and thus are closed under the standard Boolean operations of conjunction, disjunction, and negation.

By formulating the Truthmaker Principle as (TM), Lewis endorses what Armstrong calls "Truthmaker Necessitarianism": if *T* is a truthmaker for *P*, then there is no possible circumstance in which

T exists but *P* fails to be true. For to allow that there might be such a circumstance is just to concede that the existence of *T*, by itself, is not enough to make *P* true, that something more is needed (see Armstrong 2004, 6–7). Truthmaking then, unlike, for example, causation, isn't a merely contingent affair.

Formulating the Truthmaker Principle as (TM) commits Lewis also to the converse of Truthmaker Necessitarianism: if the existence of *T* strictly implies *P*, then *T* is a truthmaker for *P*. This consequence of (TM) appears to be problematic. For example, supposing arithmetical truths to be necessary, it makes me a truthmaker for the proposition that $2 + 2 = 4$. More generally, it makes any (existent) entity whatsoever a truthmaker for any necessary truth. Truthmaking for necessary truths is thus trivialized.

Lewis has little to say about the problem of truthmaking for necessary truths. I take it that there are three main lines of response. First, one could simply accept that necessary truths lack non-trivial content: they apply indiscriminately to everything, and so everything is on a par with respect to making them true. This response might be acceptable for logical truths, and perhaps even for analytic truths; but it seems especially implausible for truths of mathematics or metaphysics. Second, one could replace strict implication with a more discriminating *relevant* entailment relation: my existence does not *relevantly* entail that $2 + 2 = 4$.[8] Third, one could introduce alongside the metaphysically possible worlds *mathematically* possible worlds wherein the mathematical entities reside. (This is the response that I would prefer.) Mathematical truths are, strictly speaking, true only in mathematical worlds, and so not, strictly speaking, necessary. Since I do not exist in any of the mathematically possible worlds, I am not a truthmaker for any (purely) mathematical truth. Only mathematical entities, on this approach, can be truthmakers for mathematical truths. Nor are we forced to say that any mathematical entity is a truthmaker for any mathematical truth. The number 1, for example, is arguably not a truthmaker for $2 + 2 = 4$ because there are mathematical worlds in which mod 3 arithmetic holds, and so in which the number 1 exists but the number 4 does not, and so $2 + 2 \neq 4$. Although Lewis suggests briefly that he would favor some version of the second (relevant entailment) response, he chooses instead in all of his discussions of truthmaking to sidestep the issue by considering only a restricted version of (TM) according to which only *contingent* propositions are required to have truthmakers. For the remainder of this essay, I will suppose that (TM) (and any variant introduced below) has been restricted in this way, and I will say no more about the problem of finding truthmakers for necessary truths.[9]

Next, I turn to the task of finding truthmakers for various sorts of truth. I start with cases of truthmaking that are undisputed common ground, acceptable to both a thing theorist and a states-of-affairs theorist. Consider first *essential predications*. Suppose, for example, that the property of being a dog is essential to whatever has it, and that Fido is some actual dog. Then, Fido is a truthmaker for the proposition that Fido is a dog. For Fido is such that, necessarily, if he exists then he is a dog. Similarly, Fido is a truthmaker for the existential proposition that dogs exist. For Fido is such that, necessarily, if he exists then dogs exist. Now consider any proposition that is strictly implied by *Fido is a dog*, for example, *Fido is a mammal*. Clearly, Fido is also a truthmaker for this weaker proposition. More generally, we have as a consequence of (TM) what Armstrong (2004, 10) calls *the Entailment Principle*, but with strict implication standing in for entailment:

> **(EP)** Whenever *T* is a truthmaker for a proposition *P*, and *P* strictly implies *Q*, then *T* is also a truthmaker for *Q*.

It follows from (EP) that the truthmaking relation is one–many: a single entity is a truthmaker for many – infinitely many – true propositions.

Now, let's consider truthmakers for disjunctions and conjunctions. Suppose that the property of being a cat is essential to whatever has it, and that Tabby is some actual cat. Consider the disjunctive proposition that Fido is a dog or Tabby is a cat. Since a disjunction is strictly implied by each of its disjuncts, it follows immediately from (EP) that both Fido and Tabby individually are truthmakers for the proposition that Fido is a dog or Tabby is a cat. The truthmaking relation, then, is many–one as well as one–many. In particular, disjunctive entities needn't be introduced to make disjunctions true. Now, what about the conjunctive proposition that Fido is a dog and Tabby is a cat? Neither Fido nor Tabby alone can make the conjunction true, since there are worlds where Fido exists without Tabby, and vice versa. The most natural thing to say, surely, is that *they together* make it true, that in this case the truthmaking relation holds between *two* things and a proposition, and that, in general, the truthmaking relation takes a plural argument in its first place. This suggests that we consider a plural version of the Truthmaker Principle:

(TMP) For every true proposition P, there exist some one or more entities T_1, T_2, . . . such that, necessarily, if all of the *T*s exist, then P is true. (cf. Lewis 2001a, 607)

Is (TMP) a weaker demand for truthmakers than (TM)? That depends on whether, whenever some entities plurally make a proposition true, there is always a single entity that makes the proposition true on its own. There are two prominent candidates for bridging the gap between (TMP) and (TM): classes and mereological sums. Suppose first that one is a realist about classes. Then classes, it seems, can bridge the gap, since, first, class formation is universal – whenever there are some things, there is a class having as members all and only those things – and, second, classes have their members essentially. It follows that whenever T_1, T_2, . . . are plurally truthmakers for P, the class of T_1, T_2, . . . is a single truthmaker for P.[10] But realism about classes is controversial, and if instead one accepts a fictionalist or structuralist account of classes, classes lack the bona fides to perform the task of truthmaking.

I suggest, then, that we look elsewhere for bridging the gap. So suppose now that one is a realist about the mereological sums of classical mereology. In particular, assume (with Lewis and Armstrong) universal composition: whenever there are some things, there is a mereological sum of those things. Assume further mereological essentialism according to which wholes have their parts essentially: for any thing and any part of that thing, necessarily, if the thing exists, then the part exists as well. Then sums can bridge the gap between (TMP) and (TM). Whenever there are *T*s such that, necessarily, if all of the *T*s exist, then P is true, then the sum of the *T*s is such that, necessarily, if the sum exists, then P is true. Mereological essentialism, of course, is controversial in ordinary contexts: my left hand is part of my body, but my body, we ordinarily think, could have existed though it never had a left hand as a part. A counterpart theorist such as Lewis, however, will allow that there are also contexts that evoke a counterpart relation that makes mereological essentialism true. If truthmaking contexts evoke such a counterpart relation, then we have as much of mereological essentialism as we need. (See section 11.5 for further discussion of mereological essentialism.) In any case, since nothing that follows hangs on the difference between (TMP) and (TM), I will forgo neutrality by assuming universal composition and mereological essentialism (in truthmaking contexts), and focusing henceforth only on (TM). Taking sums of truthmakers to be truthmakers allows that truthmakers need not be *ordinary* things; for example, a truthmaker for the conjunction, Fido is a dog and Tabby is a cat, is the sum of Fido and Tabby. But, ordinary or not, nothing beyond the category of things has had to be introduced to serve as truthmakers for the propositions thus far considered.

Once we accept universal composition and mereological essentialism, we can establish a flipside to the Entailment Principle that I will call the Parthood Principle:

(PP) If T is a truthmaker for P, and T is a part of T', then T' is a truthmaker for P.

Call the mereological sum of all (actual, existing) entities the *world*. It follows immediately from (PP) that, if a proposition has any truthmakers at all, then the world is a truthmaker for the proposition. Say that an entity *discerns* one proposition from another just in case it is a truthmaker for the one but not the other. Then, the world is the least discerning truthmaker, and, as such, not a very interesting truthmaker.[11] If, as Armstrong thinks, the search for truthmakers is a rival to Quine's method for uncovering the ontological commitments of propositions, we should be searching for more discerning truthmakers; it is no news, after all, that the world exists. Two ways of supplementing the Truthmaker Principle so as to require more discerning truthmakers come readily to mind. First, atomic truths (and intrinsic predications generally) should have discerning truthmakers:

(SUP$_1$) For any atomic truth $Ra_1a_2\ldots$,there exists a truthmaker involving at most a_1, a_2, $\ldots$ and R.

For the states-of-affairs theorist, of course, the corresponding atomic state of affairs is the required discerning truthmaker. Second, although in general the truthmaking relation is many-many, for the case of atomic propositions, there are truthmakers with respect to which the relation is one–one:

(SUP$_2$) Distinct atomic truths have distinct truthmakers.

Again, this supplement to the Truthmaker Principle is automatic for a states-of-affairs theorist: atomic states of affairs are the required discerning truthmakers. The extent to which a thing theorist can accept these supplements to the Truthmaker Principle will be discussed in section 11.5.

Let's take stock. Thus far, we have found truthmakers for essential (monadic) predications, for singular existential propositions, for general existential propositions that generalize over essential properties, and for propositions that can be generated from these by taking disjunctions and conjunctions. And these truthmakers have all been compatible with a thing ontology, where mereological sums of things are taken also to be things. But what about all the propositions not yet included? Don't they too have to be ontologically grounded? The acceptance of (TM), what Armstrong calls *Truthmaker Maximalism*, demands that *every* truth have a truthmaker. The Truthmaker Maximalist needs, then, to find truthmakers for inessential predications, both monadic and polyadic, for general propositions, and for negative propositions. According to Armstrong, a thing ontology will not be able to meet this demand. Finding appropriate truthmakers for these propositions leads to an ontology of states of affairs.

11.2 Truthmaking and States of Affairs

I turn now to Armstrong's theory of truthmaking and the introduction of states of affairs. I start with the simplest case: a predication where the property predicated is intrinsic but not essential to the subject. For example, consider the (singular) proposition I express when, looking at a red ball, I say: "the ball is red." Is the ball a truthmaker for this proposition? Although it may seem as though, in some sense, the ball does indeed make this proposition true (see later in this section, and also section 11.5), taking the ball to be a truthmaker is incompatible with (TM) given our ordinary attributions of modality *de re*. For the ball is only contingently red; it could have existed and been blue, or some other color. So, it is false that the ball is such that, necessarily, if the ball exists, then the ball is red; the ball fails as a truthmaker. Somehow, a truthmaker for the proposition that the ball is red must carry the redness with it wherever it goes. Suppose, then, that there is an immanent universal of redness that the ball instantiates; and consider the mereological sum of the ball and this universal.

Could this sum be a truthmaker for the proposition that the ball is red? Although the sum necessitates that the ball exists and that redness exists (is somewhere instantiated), it does not necessitate that the ball is red; the sum could exist if the ball were blue and something else were red. (The class having as members the ball and the universal does no better, for a similar reason.) A truthmaker for the proposition that the ball is red must somehow *unify* the ball and redness, and the unification must itself be necessary. The "obvious candidate" for such a unified truthmaker, Armstrong claims (1997, 116), is the *state of affairs* of the ball's being red. For this state of affairs is such that, whenever it exists, not only do the ball and redness both exist, but the ball instantiates redness. States of affairs, if such entities exist, are made for the role of truthmaking. Accepting the demand for truthmakers, then, gives good and sufficient reason for believing that states of affairs exist.

Armstrong calls the above argument for states of affairs *the truthmaker argument*, and says (in Armstrong 1997, 115) that it is "perhaps the fundamental argument of this book." If the argument is sound, then the demand for truthmakers has substantial ontological consequences. Three caveats, however, are in order. First, positing a state of affairs to serve as truthmaker for an intrinsic predication is only justified, given Armstrong's sparse account of states of affairs, when the property predicated is (or corresponds to) a universal. Without that restriction, the truthmaker argument could be used to support the existence of negative, or disjunctive, states of affairs. Second, as Armstrong concedes, states of affairs might not be the only entities that can fill the truthmaking role (Armstrong 1997, 119). Indeed, if we switch from an ontology of universals to an ontology of tropes (particularized properties), and we allow that tropes are *non-transferrable* – that a trope cannot be instantiated by anything other than what actually instantiates it – then tropes can take the place of states of affairs as truthmakers for intrinsic predications. But, for lack of space, I will say no more in this essay about tropes as truthmakers. Problems for taking states of affairs to be truthmakers tend to have parallel problems for non-transferrable tropes; and, in any case, Lewis's critique of (TM) was directed predominantly at Armstrong's account in terms of states of affairs.

Third, Armstrong identifies what he calls "thick" particulars with certain conjunctive states of affairs, which has the effect of allowing thick particulars to be truthmakers for intrinsic predications. Consider a state of affairs of *a's being F*. This state of affairs has two constituents: the "thin" particular, which I will call *a*−, and the universal *F*; the relation of instantiation holds between *a*− and *F*. The thin particular, according to Armstrong, is "the particular in abstraction from its properties". (Armstrong 1997, 123) The thick particular, which I will call *a*+, on the other hand, is "the particular taken along with all and only the particular's non-relational properties" (Armstrong 1997, 124).[12] Let *N* (for nature) be the conjunction of all the universals instantiated by *a*-. Then, the thick particular, *a*+, according to Armstrong, is the state of affairs of *a's being N*. Now, back to our red ball. The proposition that the ball is red predicates a property of a thin particular. In the truthmaker argument, it is this thin particular that fails to be a truthmaker for the proposition. But if instead we consider the ball as a thick particular, a conjunctive state of affairs, then the ball *is* a truthmaker for the proposition after all; for the state of affairs of the ball being red is a conjunct (and so for Armstrong a part) of the thick particular, and must exist whenever the thick particular exists.[13] This concession undermines the claim, from the perspective of a states-of-affairs theorist, that a thing ontology is inadequate to the task of providing truthmakers for intrinsic predications; for it is thick particulars, presumably, not thin particulars, that a states-of-affairs theorist should take to be the things of the thing theorist. (In section 11.5, we will see that the claim is also undermined from the perspective of a thing theorist who accepts counterpart theory.) At best, the states-of-affairs theorist can claim that things cannot provide sufficiently discerning truthmakers for intrinsic predications, since the ball is a truthmaker not only for the proposition that the ball is red, but also for other intrinsic predications, such as the proposition that the ball is round. But put this aside for now. Lewis (prior to Lewis 2003) accepts the truthmaker argument, that the demand

for truthmakers leads to the postulation of states of affairs; his target, rather, is (TM) and the states of affairs themselves.

In any case, perhaps the Truthmaker Principle makes a stronger case for states of affairs when applied to general or negative propositions. Consider, for example, the general proposition that all humans weigh less than a ton. (Pretend that *being human* and *weighing less than a ton* are universals.) This proposition, if true, is contingently true: it is not impossible for there to exist a supersized human weighing more than a ton. But no thing (or thick particular) could be a truthmaker for the proposition because, assuming even a weak Humean principle of recombination, no thing necessarily excludes the existence elsewhere of a supersized human. In general, no contingent general proposition is made true by a thing, because any thing is compatible with the existence of a distinct thing that is a counterexample to the general proposition.[14]

To meet the demand for truthmakers for general propositions, Armstrong introduces *totality states of affairs*. Say that a sum *totals* a property if and only if everything that has the property is a part of the sum. For Armstrong, the totalling relation is a dyadic universal, second-order in its second *relatum*. Whenever a sum stands in the totalling relation to a property, a (second-order) *totality state of affairs* exists.[15] For example, there is a totality state of affairs asserting that the sum of all humans totals the property *being human*. Now return to the proposition that all humans weigh less than a ton. For each human, consider the state of affairs asserting that he or she is human, and the state of affairs asserting what that human weighs. Conjoin all these states of affairs with the above-mentioned totality state of affairs. This conjunctive state of affairs is such that, necessarily, if it exists, then all humans weigh less than a ton. We have found a truthmaker for the general proposition. More generally, consider the proposition that all *F*s are *G*s, where *F* and *G* are universals. Let $a_1, a_2, \ldots$ be all the *F*s. Then, the state of affairs of *a_1's being F & a_1's being G & a_2's being F & a_2's being G & . . . & the sum of $a_1, a_2, \ldots$ totaling F* is the sought-after truthmaker. Further, if we take $a_1, a_2, \ldots$ to be thick rather than thin particulars, then all the conjuncts are included in the one totality state of affairs, and the totality state of affairs, by itself, is a truthmaker for the general proposition.[16]

One might hope that a solution to the problem of finding truthmakers for general propositions would immediately carry over to the problem of negative existentials, since every negative existential proposition is logically equivalent to a general proposition. But if the equivalent general proposition generalizes over a negative property, the solution won't carry over unless we already have a solution to the problem of finding truthmakers for negative predications. Consider, for example, the negative existential that there are no purple swans. (Pretend that *being a swan* and *being purple* are universals.) This is equivalent to the general proposition that all swans are not purple. We have the totality state of affairs of *the sum of $s_1, s_2, \ldots$ totaling the property being a swan*, where $s_1, s_2, \ldots$ are all the swans. But we don't yet have truthmakers for the negative predications *s_1 is not purple*, *s_2 is not purple*, and so on. Moreover, taking $s_1, s_2, \ldots$ to be thick particulars won't guarantee that the totality state of affairs is a truthmaker for the negative existential unless we have a guarantee that *being purple* couldn't possibly be added to the nature of any of the *s*s.

Let us then consider the problem of finding truthmakers for negative predications such as *a is not F*, where *F* is a universal. We could, of course, simply postulate negative states of affairs to play the truthmaking role. But, if the states-of-affairs theorist cares about ontological economy, this should be a last resort. Sometimes, at least, positive atomic states of affairs can serve as truthmakers for negative predications. For example, returning to the red ball, the proposition that the ball is not blue is plausibly made true by the state of affairs of *the ball being red*. For, arguably, *being blue* and *being red* are necessarily incompatible properties. In general, if a thing *a* instantiates a determinate *F* from a determinable, then (arguably) the atomic state of affairs of *a being F* is a truthmaker for the negative predication that *a* is not *G*, where *G* is any other determinate from that determinable. But this solution (called the "incompatibility solution" by Armstrong) doesn't generalize. For it does not

seem necessary that, whenever a particular lacks a property, there must be some other property that it has that is incompatible with the property it lacks. Suppose, for example, that neutrinos can be *massless* in the sense of instantiating no mass property whatsoever, not even the property of having zero mass (if such there be). Consider the negative predication that some neutrino *n* doesn't have mass 1 kg. In this case, all positive states of affairs involving *n* are compatible with *n* having a mass of 1 kg (given a plausible principle of recombination for properties), and so cannot serve as truthmakers for the negative predication.

But perhaps totality states of affairs, which are needed in any case to serve as truthmakers for general propositions, can do double duty as truthmakers for negative predications. To see how, note that the proposition that *a* is not *F* (for universal *F*) is equivalent to a general proposition quantifying over universals: all universals instantiated by *a* are distinct from *F*. Suppose that G_1, G_2, . . . are all the universals instantiated by *a*, and that *F* is distinct from all of the *G*s. Consider the totality state of affairs asserting that the sum of G_1, G_2, . . . totals the property of being a universal instantiated by *a*. This is a truthmaker for the proposition that *a* is not *F*. For this totality state of affairs is such that, necessarily, if it exists, then no universal other than G_1, G_2, . . . is instantiated by *a*, and so (since *F* is *necessarily* distinct from all of the *G*s) *F* is not instantiated by *a*. Taking this totality state of affairs to provide the ontological ground for a simple negative predication might seem suspicious: unlike the second-order totality states of affairs postulated to ground general truths, this totality state of affairs is effectively third-order, relating a sum of properties to a property of properties. Such is the price for repudiating negative atomic states of affairs.

By introducing totality states of affairs, Armstrong provides truthmakers for general and negative propositions, where the properties generalized or negated are universals. But what about the rest of the propositions, including the vast majority of propositions that we think or assert? If we drop any concern with finding discerning truthmakers, we can retreat to the idea that the world *in toto* is a truthmaker for all truths. What is the world for a states-of-affairs theorist? The world must be taken to include, in addition to all the atomic states of affairs, the grand totality state of affairs asserting that these are all the atomic states of affairs. If the world instead were taken to be a big thing, the cosmos, it would not be a truthmaker for general and negative truths, since it would not include the *limits* that these general and negative truths depend on. Totality states of affairs are needed – at least the one grand totality state of affairs – to provide these limits, to provide the ground for general and negative truths. Or so says Armstrong, who, following Russell, rejects as inadequate the Tractarian account according to which all states of affairs are atomic and first-order.

11.3 Truthmaking and Theories of Truth

Before turning to Lewis's critique of Armstrong's truthmaking theory, it is worth considering a side dispute over the relation between the Truthmaker Principle and theories of truth. Truthmaker theorists such as Armstrong often claim that the Truthmaker Principle is a stripped-down version of the correspondence theory of truth. "Stripped-down," because the correspondence it invokes between truths and states of affairs (aka facts) is many–many, not one–one, and because it is neutral as to whether the internal structure of propositions, if any, "pictures" the internal structure of states of affairs. This allows the correspondence theorist to embrace a sparse account of states of affairs, paralleling a sparse account of the universals that are constituents of states of affairs. Just as there are no negative or disjunctive universals, so there are no negative or disjunctive states of affairs. Nonetheless, the essential core of a correspondence theory is upheld: a proposition is true if and only if it (appropriately) corresponds to some state of affairs, where the truthmaker relation, it is claimed, provides the appropriate correspondence.

Lewis disagrees. In "Forget about the Correspondence Theory of Truth" (Lewis 2001b), he argues that Armstrong's truthmaker theory is not correctly described as a version of the correspondence theory of truth, or as a rival to the other standardly mentioned theories of truth such as the redundancy, coherence, or pragmatic theories. Indeed, Lewis thinks that nothing that has gone by the name "correspondence theory of truth" should be counted among the rival theories of truth. Talk of the correspondence theory of truth should be banished from philosophy.

First, some stage setting. Lewis, himself, is a redundancy theorist, at least with respect to *propositional* truth.[17] For each proposition, there is what he calls a redundancy biconditional. For example, for the proposition that cats purr, there is the biconditional: the proposition that cats purr is true iff cats purr. This, and any other, redundancy biconditional is "trivial, necessary, and knowable *a priori*" (Lewis 2001a, 602). That is the positive side of the redundancy theory. But there is also a negative side: the redundancy biconditionals are all there is to a theory of (propositional) truth. No substantial property of truth is needed to play any theoretical role. Rather, the truth predicate is needed in ordinary language only to play a practical role: it allows us to form generalizations that "make a long story short." To use his example, the generalization "whatever the Party says is true" is equivalent to an infinite bundle of conditionals: "if the Party says that two and two make five, then two and two make five"; "if the Party says that we have always been at war with Eastasia, then we have always been at war with Eastasia"; and so on, with a conditional for every proposition. If we had world enough and time, perhaps we could assert each member of the bundle individually; but we don't, and hence the need for a device, such as the truth predicate, which allows us to assert the entire bundle with a single, compact sentence. But the important point is this: since this bundle of conditionals has nothing especially to do with truth, it follows that the generalization also has nothing especially to do with truth. The truth predicate is here merely a syntactic device for increasing the expressive power of our finitary language. (Alternatively, we could have introduced propositional quantifiers and pro-sentences. But we didn't, not in ordinary language.)

How does the redundancy theory relate to the other standard theories of truth? Most of these theories, it seems, must reject the positive side of the redundancy theory. For example, a version of the pragmatic theory of truth holds that biconditionals such as the following are *a priori*: the proposition that cats purr is true iff it is useful to believe that cats purr. If the redundancy biconditionals were also *a priori*, it would follow that the biconditional – cats purr iff it is useful to believe that cats purr – was itself *a priori*, which, Lewis says, it manifestly is not. Putative correspondence theories of truth, on the other hand, can accept that the redundancy biconditionals are *a priori* (and Lewis assumes throughout that they do). They conflict with the redundancy theory instead only by rejecting its negative side, by denying that the redundancy biconditionals are all there is to truth.

Now, back to the main argument. For a putative theory of truth to be aptly termed a "correspondence theory" and be counted among the rival theories of truth, according to Lewis, it would have to meet at least the following four conditions: first, it would have to go beyond the redundancy theory, thus conflicting with its negative side; second, it would have to conflict with the positive doctrines of the other theories of truth, such as the coherence and pragmatic theories; third, it would have to be aptly summarized by the slogan "truth is correspondence to fact"; and fourth, it would have to be a theory of truth, not a bundle of claims having nothing especially to do with truth. Lewis conjectures that all putative correspondence theories will fall into one of two camps. Those in the first camp, in effect, identify "facts" with true propositions. These theories are vacuous: it is no news to be told that truths correspond to true propositions (if truths themselves are propositions, as we are supposing). These putative correspondence theories, then, fail to satisfy the first two conditions: they do not go beyond the redundancy theory, nor do they conflict with the coherence or pragmatic theories. "Away with them!" Lewis cries.

The other sort of putative correspondence theory is represented by Armstrong's Truthmaker Principle. On this sort of theory, facts are not identified with true propositions, but with "Tractarian" facts, Armstrong's states of affairs. This theory is certainly not vacuous, and satisfies the first two conditions. But, according to Lewis, it fails to satisfy the third and fourth condition. It fails to satisfy the third condition, Lewis claims, because truthmakers need not be states of affairs, and so the theory isn't aptly summarized by the slogan that "truth is correspondence to fact." For example, as we saw in section 11.1, *any* entity is a truthmaker for the proposition that that entity exists, and that entities of its (essential) kind exist. This objection, however, is not in my view very serious. Armstrong presumably holds that any entity whatsoever, if not itself a state of affairs, is a constituent of some state of affairs. Moreover, any state of affairs is a truthmaker for the existence of any of its constituents. It follows from the transitivity of the truthmaking relation that, whenever an entity makes a proposition true, some state of affairs also makes that proposition true. Thus, all truths have states of affairs among their truthmakers, and the slogan seems to be captured well enough, even if some truthmakers are not states of affairs.

The chief reason to deny that Armstrong's theory should be called "a correspondence theory of truth" is that it fails to satisfy the fourth condition: it is a bundle of claims having nothing especially to do with truth, and therefore is not really a theory of truth at all. According to Lewis, the word 'truth' occurs in the Truthmaker Principle just for the purpose of making a long story short. The Truthmaker Principle is equivalent to a bundle of biconditionals, one member of which, for example is the following: the proposition that cats purr is true iff there exists some entity T such that, necessarily, if T exists, then cats purr. But, given the redundancy biconditionals, this biconditional is equivalent to: cats purr iff there exists some entity T such that, necessarily, if T exists, then cats purr. This biconditional says nothing of truth; rather, it claims that the purring of cats is *existentially grounded*. Similarly, the other biconditionals of the bundle make claims about the existential grounding of all manner of things – the flying of pigs, or what-have-you – but say nothing about the concept of truth. Lewis concludes, "if the Truthmaker Principle amounts to a bundle of claims that are not at all about truth, it should not be called a 'theory of truth' " (Lewis 2001b, 279). Consequently, disputes over whether or not to accept the Truthmaker Principle should be sharply separated from disputes over what is the correct theory of truth.[18]

11.4 Lewis's Critique of (TM) and a States-of-Affairs Ontology

I turn finally to Lewis's critique of Armstrong's truthmaker theory.[19] Following Lewis, I focus in this section on just two types of proposition: negative existentials and inessential (monadic) predications. Offhand, Lewis claims, we do not expect such propositions to have truthmakers. Consider, for example, the proposition that there are no unicorns. Intuitively, this is true not because of anything that exists, but because of what fails to exist. Sure, we can say: the *absence of unicorns* makes the proposition true. But only someone beholden to a naive theory of reference would take this to imply that absences populate the world, along with people, and planets, and protons. Consider now the proposition that the ball is red. (Pretend, as before, that *being red* is a universal.) Intuitively, this proposition is true not because of *what* things there are, but because of *how* things are. Sure, we can form a gerundial phrase, 'the ball's being red'; and using that phrase, we can say that such-and-such is true in virtue of the ball's being red. But such use of the gerundial phrase, by itself, should not lead us to say that the world is populated by an entity, *the ball's being red*, distinct from, but co-located with, the ball.

To put flesh on these intuitive bones, Lewis recasts the Truthmaker Principle as a principle as to how possible worlds must differ. Before recasting, however, we first need to strengthen the principle. Truthmaker theorists typically hold that (TM) is *necessarily* true: contingently false propositions *would*

have had truthmakers had they been true.[20] Appending 'necessarily' to the front of (TM) and regimenting the result in the language of possible worlds gives:

> **(TM+)** For any proposition *P* and any world *W*, if *P* is true in *W*, there exists some entity *T* in *W* such that, for any world *V*, if *T* exists in *V*, then *P* is true in *V*. (Lewis 2001a, 606)

(TM+) implies: For any two worlds *W* and *V* and any proposition *P*, if *P* is true in *W* but not in *V*, then there exists some entity *T* in *W* that does not exist in *V*. For if *T* is a truthmaker for *P* in *W*, then *T* cannot exist in *V* lest *P* be true in *V*. But on the abundant conception of propositions being assumed, for any two worlds *W* and *V*, there is a proposition true in *W* but not in *V*: the proposition that *W* is actualized. The following difference-making principle therefore follows from (TM+):

> **(DM)** For any two worlds *W* and *V*, there exists some entity *T* in *W* that does not exist in *V*. (Lewis 2001a, 606)

(DM) is a *two-way* difference-making principle: for any two worlds, each world contains some entity not contained in the other.[21]

The case of negative existentials, however, suggests that the difference between world populations need not be two-way. There are no unicorns in our world, but there are unicorns in some other possible worlds (I suppose). In moving from our no-unicorn world to a world populated with unicorns, why can't we simply *add* unicorns to the population? Why must we also take something away?

The case of (inessential) predications suggests that there need not even be a one-way difference in population. Consider a red ball and green bat in our world, and another world where that ball and bat have switched colors: the ball in the other world has the exact shade of green that the bat has in our world; the bat in the other world has the exact shade of red that the ball has in ours. Why must there be any difference in population between these two worlds? Why must a change in *how* things are bring with it a change in *what* things there are?

These difference-making considerations against the Truthmaker Principle are suggestive, but they are not likely to have much force for a states-of-affairs theorist such as Armstrong. They presuppose that the right way to think of a world population is in terms of things, or things and universals, not in terms of states of affairs. To get a more decisive argument against (TM+), Lewis attacks directly the states of affairs that (he allows) would be needed to serve as truthmakers. States of affairs violate two principles that are fundamental to Lewis's metaphysics: the uniqueness of composition, and the Humean denial of necessary connections. Because of these violations, Lewis concludes that an ontology of states of affairs is "bad news for systematic metaphysics" (2001a, 611).

Let's start with the composition of states of affairs. Lewis holds to a twofold principle of uniqueness of composition: "there is only one mode of composition; and it is such that, for given parts, only one whole is composed of them" (Lewis 1986c, 92). For Lewis, the laws of this single mode of composition are given by classical mereology. Many philosophers reject uniqueness of composition because they accept sets into their ontology, and hold that sets are composed of the individuals that are their members, or their members' members, or.... On this view, distinct sets, such as {a, {b}} and {{a}, b} are composed of the same elements. Lewis argues, however, that the formation of sets involves two operations: the forming of singletons and the forming of unions. Only the latter operation is a mode of composition, and the composition is mereological. Sets, then, on Lewis's account, do not violate uniqueness of composition.[22]

Now, consider states of affairs *a's being F*, *b's being G*, and *a's having R to b*, where *a* and *b* are particulars and *F*, *G*, and *R* are universals. Somehow, the state of affairs of *a's being F* is supposed to be composed of the particular *a* and the universal *F*. But, as noted in §2, the composition cannot be

mereological because (on standard essentialist assumptions) the sum $a + F$ can exist even though the state of affairs of *a's being F* does not. Moreover, whatever sort of composition is involved, it is not unique. The dyadic states of affairs of *a's having R to b* and *b's having R to a* are distinct (at least if *R* is not necessarily symmetric), even though they are composed of the same particulars and universals. And, for good measure, the conjunctive states of affairs of *a's being F & b's being G* and *b's being F & a's being G* are distinct, even though again they do not differ in their (ultimate) components. For Lewis, the idea that states of affairs can be unmereologically composed in this way from particulars and universals makes them totally mysterious.

Consider next how states of affairs violate the Humean denial of necessary connections. To capture the Humean prohibition, Lewis introduces a *principle of recombination*, initially formulated as follows: "anything can coexist with anything else, and anything can fail to coexist with anything else" (Lewis 1986a, 88). The first half, strictly speaking, is a prohibition against necessary *exclusions*. Lewis's illustration: if there could be a unicorn, and there could be a dragon, then there could be a unicorn and a dragon side by side. How should this be interpreted in terms of worlds? Since worlds do not overlap for Lewis, a unicorn from one world and a dragon from another cannot *themselves* exist side by side. The principle is to be interpreted in terms of *intrinsic duplicates*: in some world, a duplicate of the unicorn and a duplicate of the dragon exist side by side. The second half of the principle of recombination is the prohibition against necessary *connections*. Spelled out in terms of worlds and duplicates, it says: whenever two distinct things coexist in a world, there is another world in which a duplicate of one exists without a duplicate of the other.[23] Lewis's illustration: since a talking head exists contiguous to a living human body, there could exist an unattached talking head, separate from any living body. More precisely: there is a world at which a duplicate of the talking head exists but at which no duplicate of the rest of the living body exists.

Now, suppose that the atomic proposition that *a* is *F* is true for some particular *a* and universal *F*, and consider the atomic state of affairs, *a's being F*, that according to the states-of-affairs theorist makes it true. As noted above, *a's being F* is not mereologically composed of *a* and *F*, and thus *a's being F* and *a* are mereologically distinct. Moreover, since *a's being F* is a truthmaker for the proposition that *a* is *F*, *a's being F* is necessarily such that if it exists, than *a* is *F*. But also, necessarily, if *a* is *F*, then *a* exists. Thus *a's being F* cannot fail to coexist with *a*, and the prohibition against necessary connections is violated.

Next, suppose that the negative existential proposition that there are no *F*s is true, and consider a state of affairs *S* that according to the states-of-affairs theorist makes it true. (For Armstrong, *S* is, or includes, a totality state of affairs.) Since *S* is a truthmaker for the proposition that there are no *F*s, *S* cannot possibly coexist with an *F*. Thus, Lewis concludes, the prohibition against necessary exclusions is violated.

Lewis's second argument against states of affairs, however, falls short of the mark, as can be seen by recasting it in terms of worlds and duplicates. *S*, we are supposing, exists in the actual world where there are no *F*s. Let *W* be a world where *F*s exist, and let *a* be an *F* that exists in *W*. And suppose, as is usual, that *F*, being a universal, is an intrinsic property of *a*. The most that Lewis's argument establishes is that *S itself* cannot coexist with a duplicate of *a*: there is no world in which both *S* and a duplicate of *a* exist. But what needs to be shown is that there is no world in which *a duplicate of S* and a duplicate of *a* both exist. If the exclusionary power of *S* arises from its *extrinsic* nature – as one would naturally suppose – then the argument fails. The way in which *S* manages to necessarily exclude *a* may be mysterious – Lewis no doubt thinks that it is – but not because it is a violation of Lewis's Principle of Recombination.

The first argument against states of affairs also faces difficulties. Although it establishes a violation of the Principle of Recombination, as Lewis understands it, it has no force against a states-of-affairs theorist who thinks that states of affairs have an unmereological mode of composition. For such a

theorist, the state of affairs of *a's being F* and the particular *a*, though *mereologically* distinct, are not distinct *simpliciter* because *a* is an unmereological component of *a's being F*. And since they are not distinct *simpliciter*, necessary connections between them are excusable, indeed, are to be expected. The Humean denial of necessary connections between "distinct" existences does not apply. So Lewis's argument that states of affairs violate the Humean denial of necessary connections depends upon his defense of the principle of Uniqueness of Composition. Lewis's metaphysical views on composition are primary.

Interestingly, Lewis thinks it is the other way around, claiming that the complaint involving necessary connections "subsumes" the complaint involving unmereological composition (2001a, 611). He says that unmereological composition can be defined in terms of necessary connections, so if he could understand the necessary connections, then he could understand unmereological composition. Although he doesn't elaborate on this, the definitions he has in mind are, presumably, something like the following. First, some terminology. Let's say that *a* is a *part* of *b* just when *a* is a mereological component of *b*; and let's say that *a* is a *constituent* of *b* just when *a* is an unmereological component of *b*. ('Component', then, is the inclusive, neutral term.) We suppose that 'part' (or 'mereological component') is already understood. We can define the neutral notion of being a component as follows: *a* has *b* as a *component* iff, necessarily, whenever *a* exists, *b* exists. We can then define the unmereological notion of a constituent like this: *a* has *b* as a *constituent* iff *a* has *b* as a component, but *b* is not a part of *a*. Then, unmereological composition can be defined in terms of 'constituent' just as mereological composition is defined in terms of 'part'. These definitions show that if we could understand necessary connections between mereologically distinct entities, then we could understand unmereological composition. And that is why Lewis thinks the Humean objection to states of affairs subsumes the mereological objection.

This puts pressure on Lewis to defend the Humean denial of necessary connections. For all the importance that it plays throughout Lewis's metaphysics, his defense essentially comes down to this: "it is the Humean prohibition against necessary connections that gives us our best handle on the question what possibilities there are" (2001a, 611). But that does not help with determining the exact *scope* of the principle. The Humean prohibition does not apply without restriction: entities that are not mereologically distinct, that have a part in common, are allowed to stand in necessary connections. Lewis needs a principled reason why these, and only these, entities are excluded from the scope of the principle. One good reason, I think, would be this. Parthood is partial identity; so it is as much to be expected that there are necessary connections between an entity and its parts as that there are necessary connections between an entity and itself; and, surely, no entity could fail to coexist with itself! But Lewis has backed away somewhat from the view that composition is identity – he says instead that composition is *analogous* to identity (1991, 84–7) – and so it is unclear whether he can avail himself of this response. In any case, we see again that it is the account of composition that is primary, since one's views on composition inevitably inform the interpretation and defense of the Humean denial of necessary connections. This turns Lewis's claim on its head: if we could understand why mereological composition is, or is not, the only mode of composition, we could understand why the Humean prohibition always, or only sometimes, applies to mereologically distinct entities. The mereological objection subsumes the Humean objection.[24]

11.5 (TM) and a States-of-Affairs Ontology, Reconsidered

Over the course of 15 years and four papers (1986c, 1992, 1998, 2001a), Lewis never wavered in his critique of the Truthmaker Principle and the states-of-affairs ontology that goes with it. Then, toward the end of his career, he takes it all back – or so it might seem. In Lewis (2003), he argues

that intrinsic predications have truthmakers after all. In a postscript to that paper (Lewis and Rosen 2003), he argues that negative existentials have truthmakers as well. In an unpublished paper written shortly thereafter (chapter 2, this volume), he withdraws his objections to a states-of-affairs ontology. This is certainly a *new* view; but is it a *change* in view? Lewis thought not. His critique had relied upon standardly accepted attributions of essential properties; he did not want his own views in the metaphysics of modality, his own brand of essentialism, to prejudice the debate. When he stopped to consider what could be said from his own perspective, there was a radical shift. A counterpart-theoretic account of essential properties allows for a more flexible interpretation of the Truthmaker Principle, an interpretation under which it can be taken to be literally true without running afoul of Uniqueness of Composition or the Humean prohibition. For a truthmaker theorist who does not embrace counterpart theory and the inconstancy of *de re* modality, Lewis's critique still stands.

11.5.1 Truthmaking and Counterpart Theory: Lewis's Approach

First, we need some background on counterpart theory.[25] As noted in section 11.1, the Truthmaker Principle (for Lewis) is an assertion of modality *de re*. It depends for its interpretation on how a world represents *de re* of an object whether it exists in the world, and what properties it has in the world. For Lewis, since objects inhabit – are part of – only one world, this must be done by considering an object's *counterparts* in other worlds. Thus, a world *W* represents *de re* of an actual object *a* that *a* exists in *W* just in case *a* has a counterpart that inhabits *W*. A world *W* represents *de re* of *a* that *a* has property *ø* just in case some counterpart of *a* inhabits *W* and is *ø*.[26] The counterpart relation can then be used to characterize which properties of *a* are contingent, and which essential: *a* has *ø* *contingently* just in case *a* has *ø*, but some counterpart of *a* in some world doesn't have *ø*; *a* has *ø essentially* just in case every counterpart of *a* in every world has *ø*.[27]

Three conditions that must be met by any proposed counterpart relation are especially significant for what follows (see Lewis 2003, 33). First, the counterpart relation must be based on qualitative similarity. Second, the respects of similarity that count must be predominantly intrinsic. And, third, the respects of similarity that count must be "important," where what counts as important can vary, within limits, from context to context. This leads to the *inconstancy* of *de re* modality, and to a multiplicity of admissible counterpart relations. Lewis writes:

> Today, thinking of Saul Kripke as essentially the occupant of a distinguished role in contemporary philosophy, I can truly say that he might have been brought by a stork. Tomorrow, thinking of him as essentially the man who came from whatever sperm and egg he actually came from, I can truly say that he might never have had a philosophical thought in his life. I would be right both times, but relative to different, equally admissible, counterpart relations. (2003, 28)

One of the chief benefits of counterpart theory with multiple counterpart relations is that it allows us to hold on to ordinary ways of counting and individuating objects. Consider the stock example from Gibbard (1975): suppose I peer into a room that is empty except for a statue of Goliath made of clay. We ordinarily think that there is one object in the room, an object that is both a statue and a lump of clay. But philosophers often disagree. They say that the statue – let's call it Goliath – and the lump of clay – let's call it Lumpl – cannot be one and the same object. For there is a property Lumpl has that Goliath lacks: Lumpl, but not Goliath, *could have survived a squashing*. By an application of Leibniz's Law, Lumpl is not identical with Goliath. But if 'could have survived a squashing' expresses different modal properties when applied to 'Lumpl' and when applied to 'Goliath', the inference from Leibniz's Law is invalid. The counterpart theorist diagnoses the situation as follows.

There is one counterpart relation, *counterpart*$_S$, under which all counterparts of the object are statues; with respect to the counterpart$_S$ relation, the object is essentially a statue, and could not have survived a squashing. There is a different counterpart relation, *counterpart*$_L$, under which all counterparts of the object are lumps of clay, but need not be statues; with respect to the counterpart$_L$ relation, the object is not essentially a statue, and could have survived a squashing. When we use the name 'Goliath', we typically (though not invariably) evoke the counterpart$_S$ relation. So interpreted, 'Goliath could have survived a squashing' falsely attributes the property *has a counterpart*$_S$ *that survives a squashing* to the object in the room. When we use the name 'Lumpl', we typically evoke the counterpart$_L$ relation. So interpreted, 'Lumpl could have survived a squashing' truly attributes the property *has a counterpart*$_L$ *that survives a squashing* to the object. Sometimes both counterpart relations are needed for the interpretation of a single sentence, such as 'Lumpl, but not Goliath, could have survived a squashing'. The best way of capturing this in a semantic or pragmatic theory need not detain us. What matters is that counterpart theory with multiple counterpart relations allows one to attribute different modal properties to an object depending upon how the object is considered.[28]

Now, in order to apply counterpart theory to truthmaking, we need to reformulate (TM) explicitly in terms of worlds and counterparts:

> **(TMC)** For every true proposition *P*, there exists some entity *T* and some admissible counterpart relation such that, for every world *W*, if *T* has a counterpart (under that relation) in *W*, then *P* is true in *W*.[29]

(TMC) makes explicit that, for Lewis, the counterpart relation used to interpret truthmaking claims is not fixed once and for all; it may vary from context to context depending on which truthmaking claims are under discussion. I consider this further in section 11.5.2.

To illustrate (TMC), let's return to the problem of finding truthmakers for intrinsic predications such as 'the ball is red'. As noted in section 11.2, there seems to be some sense in which the ball itself can serve as truthmaker for the proposition that the ball is red; but we were prohibited from saying this because *being red* is a contingent property of the ball, and so not a part of its essence. To serve as truthmaker, the ball would have to be *essentially* red. Then, and only then, would the ball be such that, necessarily, if it exists, then it is true that the ball is red. Counterpart theory to the rescue! A flexible essentialism can allow that there are contexts with respect to which *being red* is part of the ball's essence. One way to conventionally evoke such a context is to introduce special 'qua'-names of objects, such as 'the ball qua red'. When we refer to the ball as 'the ball qua red', we evoke an unusual counterpart relation under which all counterparts of the ball are red; thus we can truly assert 'the ball qua red is essentially red'. What are these peculiar qua-entities, and how do they fit into a thing ontology? The ball qua red, of course, is nothing other than the ball itself, just as the statue Goliath is nothing other than the lump of clay Lumpl. Once again, by multiplying counterpart relations, we avoid having to multiply entities. But is there an *admissible* counterpart relation under which all counterparts of the ball are red? Suppose we start with an ordinary counterpart relation and simply add the restriction that all counterparts of the ball are red. The new counterpart relation is based on predominantly intrinsic respects of qualitative similarity, since the ordinary counterpart relation is; and the new counterpart relation is still based on respects that are important in the context, since use of 'the ball qua red' made *being red* important by suggesting that that was how we were to consider the ball. The strategy clearly generalizes. For any intrinsic predication, *a* is *ø*, we can say that *a qua ø* is a truthmaker where *a qua ø* is none other than *a* itself; for, under the counterpart relation evoked by *a qua ø*, any world in which *a* exists (by having a counterpart) is a world in which *a* is *ø* (by having a counterpart that is *ø*).

Consider next the problem of finding truthmakers for negative existentials such as the proposition that there are no unicorns. Perhaps the qua-names, and the unusual counterpart relations they evoke, can help here as well. For example, we could say that an ordinary thing, such as the Eiffel Tower, is a truthmaker for the proposition that there are no unicorns if we conceive of the Eiffel Tower as: the Eiffel Tower qua unaccompanied by unicorns. Does this qua-name evoke a counterpart relation under which all of the Eiffel Tower's counterparts are in worlds uninhabited by unicorns? Unlike the use of qua-names to show that ordinary things are truthmakers for intrinsic predications, this, according to Lewis, is a cheap trick. The supposed counterpart relation evoked by use of this qua-name is based on respects of similarity that are almost entirely extrinsic, depending neither on the intrinsic nature of the Eiffel Tower nor the intrinsic nature of its immediate surroundings. Thus, it is not an admissible counterpart relation. Even a flexible essentialism, Lewis thinks, must have limits to its flexibility. Lewis (2003) thus still denies that negative existentials have truthmakers.

But in the postscript (Lewis and Rosen 2003), Lewis reconsiders. By considering bigger truthmakers, the needed counterpart relations become less extrinsic. Although no proper part of the world can, in virtue of its intrinsic nature, make it true that there are no unicorns, perhaps the world in its entirety can do the job, where the world is just the biggest thing, the cosmos. Indeed, surely there is some sense under which the world as a whole makes it true that there are no unicorns.

A truthmaker theorist would argue, however, that no *thing* can make negative existentials true, because any thing, even the cosmos, might have been a proper part of a bigger thing. (Consider, for example, a possible world containing a series of cosmic oscillations – big bang, big crunch, big bang, and so on – and suppose one of the cycles is a duplicate of our cosmos.) If our world might have been a proper part of a bigger world, then it might have existed in its entirety while there also existed somewhere outside of its bounds a unicorn. Thus the world, if taken to be a thing, is not a truthmaker for the proposition that there are no unicorns. We can, indeed, the truthmaker theorist continues, say that the world as a whole is a truthmaker for the proposition that there are no unicorns; but only because the world is a state of affairs, not a thing, and is composed in part of totality states of affairs.

But the thing theorist who accepts the inconstancy of *de re* modality has a ready response. Although we can say that the world might have been part of a bigger world, we also can say, with no less propriety, that the world might have been bigger than it is. (For example, the world might have contained a series of cosmic oscillations.) Both of these claims are naturally and straightforwardly interpreted as modality *de re*; moreover, the very same possible world (say, with cosmic oscillations) can serve to validate both claims. To validate the first claim, we identify the world by intrinsic character alone, so that counterparts of the world must be duplicates of the world. To validate the second claim, we identify in part by extrinsic character, taking it to be essential to a world that it be a world, that is, the biggest thing there is; on this way of identifying, counterparts of the world must be *worlds*. Combining these two ways of identifying, we get a counterpart relation according to which counterparts of worlds are always duplicate worlds. This counterpart relation is admissible, being based on respects of similarity that are predominantly intrinsic. And the respect that is not intrinsic, being the biggest thing, and so unaccompanied, is clearly an important respect. To evoke this counterpart relation, we can use the qua-name: the world qua unaccompanied and intrinsicially just as it is. When the world is thus considered, the world is a truthmaker for every truth. A thing theorist, then, can endorse (TM) and Truthmaker Maximalism.

Thus far we have seen how the flexibility of counterpart theory allows *things* to be truthmakers for intrinsic predications and negative existentials; states of affairs aren't needed to satisfy (TM). But the states-of-affairs theorist is not content. Suppose that there are immanent universals, and that some particular a instantiates universals F and G. Although the states-of-affairs theorist grants that some thing – the thick particular associated with a – makes true both the proposition that a is

F and the proposition that *a* is *G*, she thinks that each of these propositions has its own, more discerning, truthmaker. The state of affairs of *a being F* makes true that *a* is *F*; the state of affairs of *a being G* makes true that *a* is *G*. To get sufficiently *discerning* truthmakers, one still needs states of affairs.

As we saw in section 11.4, Lewis had objected to states of affairs on two grounds: their unmereological composition and their involvement in necessary connections. But in an unpublished note titled "Counterparts of States of Affairs" (chapter 2, this volume) written shortly after Lewis (2003; Lewis and Rosen 2003), he withdraws these objections. With the flexibility of counterpart theory on board, states of affairs can be made innocent by identifying them with mereological sums of particulars and universals. The Truthmaker argument touted by Armstrong and previously endorsed by Lewis, that the state of affairs of *a's being F* could not be identified with the mereological sum $a + F$, rested in part on mereological essentialism, the view that the whole cannot exist without its parts, nor the parts exist without the whole. Formulated in counterpart theoretic terms, mereological essentialism amounts to this:

> **(ME)** For all entities *a* and *b*, if $a + b$ has a counterpart, $(a + b)'$, in a world *W*, then both *a* and *b* have counterparts, a' and b', in *W*, and $a' + b' = (a + b)'$; and if *a* and *b* have counterparts, a' and b', in a world *W*, then $a + b$ has a counterpart $(a + b)$ ' in *W*, and $a' + b' = (a + b)$ '.

But a flexible essentialist will say that (ME) is true in some contexts but not in others. To illustrate, consider the question whether a chair is identical with the sum of its legs, seat, and back. A dogmatic mereological essentialist must answer "no" because we can truly say that the chair could have existed without one of its legs, and that the sum could have existed even though the legs, seat, and back were never assembled into a chair. In counterpart theoretic terms: because there are counterparts of the chair that are not identical with any sum of the counterparts of the chair's parts, by (ME), the chair cannot be identical with the sum. But, once again, flexible counterpart theory comes to the rescue. There is only one thing, alternately referred to as "the chair" or "the sum of the legs, seat, and back." When we refer to it as "the chair," we create a context that evokes a counterpart relation under which the parts are not essentially tied to the whole and mereological essentialism is false. When we refer to it as "the sum of the legs, seat, and back," we create a context that evokes a counterpart relation under which the parts and the whole are essentially linked and mereological essentialism is true. One thing has different essences depending upon how it is considered.

The same treatment applies to sums and states of affairs. An enlightened states-of-affairs theorist, one who accepts a flexible essentialism, can say that the state of affairs of *a's being F* is identical with the mereological sum $a + F$. (More exactly, $a\text{-} + F$, the sum of the *thin* particular and the universal.) But this one entity, which is both a state of affairs and a sum, has different essences depending upon how it is considered. Considered as a state of affairs, it is essentially a state of affairs, and exists if and only if *a* is *F*. Considered as a sum, it is not essentially a state of affairs, and exists whenever *a* and *F* exist, whether or not *a* is *F*. Again, one multiplies essences (and counterpart relations) without multiplying entities. This same treatment applies to any state of affairs with particulars and universals as components: the state of affairs is identical with the mereological sum of those particulars and universals.[30]

Lewis's earlier objections to a state-of-affairs ontology now dissolve. The first objection was that states of affairs violate Uniqueness of Composition. But if states of affairs *are* just mereological sums, then no two states of affairs have the same components because mereological composition is unique. An opponent might, however, try to turn this on its head: don't the examples in §4 introduced to show that the composition of states of affairs is *not* unique tell against the identification of states of affairs with mereological sums? Suppose, for example, that the states of affairs of *a having R to b* and of *b having R to a* both exist, for some dyadic universal *R*. (Suppose also that *R* is not necessarily

symmetric.) Aren't these states of affairs distinct, *contra* Lewis's proposal? There are two complementary replies. First, perhaps our offhand opinion that these states of affairs are distinct comes from thinking of states of affairs as proposition-like. On that conception of states of affairs, the opinion is correct, but irrelevant. Second, even when thinking of states of affairs as Tractarian facts that compose the world, we may still be in the grip of some Leibniz's Law argument that lacks force for a flexible essentialist. In any case, notwithstanding our offhand opinions, perhaps overall theoretical considerations favor identifying these dyadic states of affairs with mereological sums, and thus with one another. Or so Lewis could claim (but see section 11.5.2 for a problem).[31]

The second objection was that states of affairs are involved in mysterious necessary connections. Whenever the state of affairs of *a being F* exists, necessarily, *a* exists and instantiates the universal *F*. Whence come these necessary connections? They arise from the counterpart relation evoked when speaking of states of affairs, and thus are no more mysterious than the workings of a flexible counterpart theory – something Lewis thinks we understand well enough. Thus, it is no mystery on this account how there could be something such that (in some contexts) it is true to say that, necessarily, it exists only if *a* is *F*. This entity, qua state of affairs, is a truthmaker for the proposition that *a* is *F*. This same entity, qua mereological sum, is not a truthmaker for the proposition that *a* is *F*. Once again, whether or not an entity is a truthmaker for a proposition depends upon how that entity is considered.

11.5.2 Truthmaking and Counterpart Theory: An Alternative Approach[32]

Lewis's strategy for rehabilitating the Truthmaker Principle, however, is open to a decisive objection. That strategy is this. For any true proposition, find some thing whose essence can be tailor-made, using flexible counterpart theory, so that the thing's existence necessitates the proposition. For *the ball is red*, use 'the ball qua red' to evoke a counterpart relation under which the ball is essentially red; for *there are no unicorns*, use 'the world qua unaccompanied by unicorns' to evoke a counterpart relation under which the world cannot coexist with a unicorn. And so on. What constrains this tailoring of essences to truths? Without constraints, we could apply the strategy according to a simple formula: for any true proposition *P* and any thing *T*, use '*T* qua inhabits a world where *P*' to evoke a counterpart relation that makes *T*'s existence necessitate *P*. Lewis would reply: '*T* qua inhabits a world where *P*' will often fail to evoke an admissible counterpart relation; *P* may be about matters that are too extrinsic, or unimportant, for *P* to be taken to be essential to *T*. I demur. I think a flexible essentialist should allow that for any *T* and *P*, *some* context could be concocted according to which the truth of *P* is essential to *T*. But put that worry aside. There is a bigger problem.

Suppose, then, that the problem of placing limits on what counts as an admissible counterpart relation is satisfactorily solved. At best, Lewis's strategy shows that, for each truth, there is a counterpart relation under which that truth has a truthmaker. What we need, I claim, is something stronger: there is a single counterpart relation under which every truth has a truthmaker. That is, we should replace (TMC) with the stronger:

> **(TMC+)** There is an admissible counterpart relation, call it $counterpart_T$, such that: for every true proposition *P*, there exists some entity *T* such that, for every world *W*, if *T* has a $counterpart_T$ in *W*, then *P* is true in *W*.

The $counterpart_T$ relation is evoked by "truthmaker contexts," contexts in which we ask whether a proposition is *made true* by some thing, or is true *in virtue of* the existence of some thing. Without a single such counterpart relation, we do not have a *uniform* interpretation of the Truthmaker Principle

under which it is true. We have only, for each instance of the Truthmaker Principle, an interpretation that makes that instance true. If we take the Truthmaker Principle to be a *schema*, an infinite bundle of assertions, one for each proposition, then a non-uniform interpretation might be good enough. But if we take the Truthmaker Principle to be a single assertion (as we have heretofore), the assertion that every truth has a truthmaker, then the variable ranging over potential truthmakers needs to be interpreted uniformly with respect to a single counterpart relation.[33] For comparison, suppose I walk into a room filled with statues made of clay and say: "some of these are essentially statues; others are essentially lumps of clay." If there are no relevant features distinguishing some of the statues from the others, I find this scarcely intelligible.

What conditions would a counterpart relation have to meet in order to validate the Truthmaker Principle? First, since any aspect of the intrinsic character of a thing may be relevant to what propositions the thing makes true, a counterpart of the thing must preserve its intrinsic character. We can guarantee this by requiring:

(C1) Counterparts$_T$ are always intrinsic duplicates.

(C1) ensures that a true intrinsic predication of a thing will have that thing as a truthmaker. Thus, if *a* has *ø*, for intrinsic *ø*, then *a* is such that, necessarily, whenever it exists it has all of its actual intrinsic properties, and so, in particular, it has *ø*.

Second, on any view that accepts the Truthmaker Principle, the world should be the supreme truthmaker, making every truth true. But, as we saw in section 11.3, if the world is a *thing* – the *biggest* thing – then the world will fail to be a truthmaker for negative existentials if the world could have been a proper part of a bigger world. The thing theorist, then, needs to require, at least in truthmaking contexts, that the world is essentially the biggest thing, that the world is essentially a world. In terms of counterparts:

(C2) Counterparts$_T$ of worlds are always worlds.

Any counterpart relation satisfying (C1) and (C2) trivially validates the Truthmaker Principle. Under any such counterpart relation, the counterpart of a world is a duplicate world. But duplicate worlds are indiscernible;[34] they agree with respect to the truth or falsity of any (qualitative) proposition. Thus, for any true proposition *P*, the world (the cosmos) is such that, necessarily, if it exists, then *P* is true. The world is a truthmaker for every truth.

If the goal were just to find a counterpart relation that makes the Truthmaker Principle true, we could stop here. But sometimes, at least, more discerning truthmakers can be had. We have seen this, so far, only for the case of intrinsic monadic predications. This is the only case that Lewis considers. But I think discerning truthmakers can be found as well for intrinsic polyadic predications, and for the negations of intrinsic predications. In each of these cases, the object or objects of predication can serve as truthmakers.

Consider first the case of negated intrinsic monadic predications. Intuitively, when asking what makes it the case that a thing has, *or fails to have*, some intrinsic property, we need look no further than the thing itself. For an intrinsic property *ø*, *a* is a truthmaker for *a is ø* if it is true that *a* is *ø*, and *a* is a truthmaker for *a is not ø* if it is true that *a* is not *ø*. In asking what constraints this puts on the counterpart relation in truthmaking contexts, there are two cases to consider: a nominalist thing theorist who rejects universals and tropes; and a realist thing theorist who accepts universals or tropes. For the nominalist thing theorist, (C1) already suffices. Suppose *a is not ø* is true. By (C1), any counterpart$_T$ of *a* in any world lacks *ø*. Thus, *a* is such that, necessarily, if it exists, then it is not *ø*. And that's what it takes for *a* to be a truthmaker for the proposition that *a* is not *ø*.

But now consider a realist thing theorist who takes things to be thick particulars composed of thin particulars and universals. In seeking a truthmaker for *a is not ø*, we need to keep track of the distinction between the thin particular, *a-*, and the thick particular, *a+*. It is the thick particular, *a+*, that the thing theorist takes to be a truthmaker for *a is not ø*. But (C1) does not suffice to guarantee that the thing *a+* is a truthmaker. For consider a world *W* where *a+* has a (unique) counterpart$_T$ *b* that is not itself a thick particular, but is instead a "middle-sized" particular properly included in a thick particular *b+*. And suppose that, although *b* does not include *ø* (since by (C1) it is a duplicate of *a+*), the thick particular *b+* does include *ø*. Now, I suppose that for any particular *a* – thin, thick, or middle-sized – it is true that *a* is *ø* iff the thin particular component of *a* instantiates *ø*. On those truth conditions, although *a+'s* counterpart$_T$ in *W*, *b*, does not include *ø*, the proposition that *a* is *ø* is nonetheless true in *W*. To defuse this counterexample, we need to require that a thick particular could not have failed to be a thick particular, that things are essentially things. In terms of counterparts:

(C3) Counterparts$_T$ of things are always things.

(C1) and (C3) together ensure that a true negation of an intrinsic predication of a thing will have that thing as a truthmaker.[35]

Finally, consider the problem of finding things to serve as truthmakers for intrinsic polyadic predications. Start with the dyadic case; the generalization to higher adicity is routine. Thus, suppose that *a* has *R* to *b*, where *R* is a fundamental relation, and thus intrinsic.[36] A flexible truthmaker theorist might hope that the sum $a + b$ will be a truthmaker for the proposition that *a* has *R* to *b*. This will require (as we saw in section 11.1) that, in truthmaking contexts, mereological essentialism (ME) holds, at least with respect to things. Does (C1) together with (ME) (for counterpart$_T$) ensure that $a + b$ is a truthmaker for *a has R to b*? Indeed, given (C1) and (ME), every counterpart$_T$ of $a + b$ is a duplicate of $a + b$, and so has a duplicate of *a* as a part that is R-related to a part that is a duplicate of *b*. But that may not be sufficient to make $a + b$ a truthmaker for *aRb*. For a counterexample, consider this. Suppose that *a* and *b* are duplicates, and that *R* is an intrinsic relation such that *aRb*, but not *bRa*. Consider a world where *a* and *b* each have a single counterpart$_T$, a' and b', respectively, and such that $b'Ra'$. By (C1), *a*, *b*, a', and b' are all duplicates of one another. By (C1) and (ME), $a' + b'$ is a duplicate of $a + b$, and so not $a'Rb'$. Then, although $a + b$ has a counterpart$_T$ in the world, no counterpart$_T$ of *a* has *R* to any counterpart$_T$ of *b*. Thus, the existence of $a + b$ does not necessitate that *a* has *R* to *b*.

The solution, as in other cases involving essential relations, is to consider not only counterparts of individuals, but also counterparts of pairs (and more generally counterparts of sequences of arbitrary length).[37] Say that two pairs, $<a,b>$ and $<a',b'>$, are *intrinsic isomorphs* iff *a* and a' are duplicates and *b* and b' are duplicates, and, for any intrinsic relation *R*, *aRb* iff $a'Rb'$. To ensure that $a + b$ will be a truthmaker for *aRb* we can, first, put the following constraint on the pair-counterpart relation:

(C4) If *a* and *b* have counterparts$_T$ in *W*, then $<a,b>$ has a counterpart$_T$ pair in *W*; and counterpart$_T$ pairs are always intrinsic isomorphs.[38]

Then, second, we modify the counterpart-theoretic semantics for doubly *de re* modal assertions so as to quantify over counterpart pairs rather than counterparts.[39] Thus, a world represents *de re* of $<a,b>$ that *a* has *R* to *b* iff some counterpart pair of $<a,b>$, $<a',b'>$, exists in *W*, and a' has *R* to b' in *W*. Given (C1)–(C4), (ME), and the modified semantics, $a + b$ is a truthmaker for *aRb* whenever *aRb* is true and *R* is intrinsic.[40]

Call any counterpart relation that satisfies (C1)–(C4) and (ME) a *truthmaking* counterpart relation. (Counterpart relations are now expanded to include sequences among their *relata*.) Truthmaking counterpart relations have at least as much claim to legitimacy as the counterpart relations admitted by Lewis. Because of (C1) and (C4), the respects of similarity that count are "predominantly intrinsic." And the extrinsic respects of similarity that count according to (C2) and (C3) – the property of being a world and the property of being a thing – are undeniably "important" in contexts where ontology is under discussion. Moreover, unlike Lewis's piecemeal approach to finding truthmakers that appeals to multiple counterpart relations, fixing on a single truthmaking counterpart relation allows for a uniform interpretation of (TM) and its supplement (SUP_1), an interpretation that makes them both literally true. ((SUP_1), recall, was the thesis: for any atomic truth $Ra_1a_2\ldots$, there exists a truthmaker involving at most a_1, $a_2\ldots$ and R.) The 'qua'-names are not needed to evoke a truthmaking counterpart relation. Rather, in truthmaking contexts, contexts in which the search for truthmakers is explicit, a truthmaking counterpart relation is automatically evoked. That provides the best explanation, for a thing theorist, as to why (TM) and (SUP_1) have the ring of truth. In considering (TM) or (SUP_1), we create a truthmaking context; and in truthmaking contexts, they *are* true.

There is one casualty, however, in the switch to a uniform interpretation of (TM): states of affairs, understood in Lewis's way as mereological sums of universals and (thin) particulars, no longer provide sufficiently discerning truthmakers to validate (SUP_2), the other supplement to (TM). ((SUP_2), recall, was the thesis: distinct atomic truths have distinct truthmakers.) The problem arises with certain relational or complex states of affairs. Consider, for example, a dyadic universal R such that aRb and bRa. By (SUP_2), since aRb and bRa are distinct atomic propositions, they have distinct (i.e. non-identical) truthmakers. A thing theorist will simply reject (SUP_2) and say that there is a single truthmaker, $a + b$, for both aRb and bRa. A states-of-affairs theorist, however, is committed to (SUP_2), and must find distinct truthmakers for aRb and bRa. But if states of affairs are mereological sums, how can the one sum, $a + R + b$, provide two distinct truthmakers? On Lewis's piecemeal approach, this is easily done: by multiplying counterpart relations, one multiplies truthmakers. The 'qua'-name '$a + R + b$ *qua a-having-R-to-b*' evokes one of these counterpart relations; '$a + R + b$ *qua b-having-R-to-a*' evokes the other. Although there is only one entity involved – $a + R + b$ – one can allow that (in truthmaking contexts) the two 'qua'-versions of this one entity count as two different truthmakers.[41] But this way of validating (SUP_2) is ruled out if we want a uniform interpretation of the truthmaking relation. If there is only a single counterpart relation with respect to which we interpret (TM) and (SUP_2), then no one entity can provide two different truthmakers. I conclude, then, that Lewis's irenic attempt to rehabilitate a states-of-affairs truthmaking theory falls short of the mark.[42]

11.6 Truth Supervenes on Being

In the previous section, we saw that a thing theorist armed with flexible counterpart theory can accept the Truthmaker Principle as literally true. That is not to say, however, that the thing theorist takes the Truthmaker Principle to be a fundamental principle of metaphysics. On the contrary, showing how to make the Truthmaker Principle true is a metaphysical sideshow, interesting only as a way of appeasing ordinary intuitions about truthmaking, and thereby placating truthmaker theorists overly enthralled by those intuitions.

Here are three reasons why the Truthmaker Principle should be no part of fundamental metaphysics. First, the Truthmaker Principle, being an assertion of modality *de re*, depends for its interpretation on context, and no fundamental metaphysical principle should be context dependent. This objection, however, can be gotten around by formulating the Truthmaker Principle in the language of possible

worlds, explicitly incorporating the truthmaking counterpart relation into the formulation. That would eliminate the context dependence.

Second, all fundamental metaphysical principles are necessary, whereas the Truthmaker Principle, even if true, is not necessarily true. Or so I claim, because I take it to be metaphysically possible that nothing exists – or, at any rate, that no contingent entity exists.[43] But the Truthmaker Principle is incompatible with such a possibility: if it is true that no (contingent) entity exists, then that very truth lacks a truthmaker, and the Truthmaker Principle is false. Admittedly, this objection lacks bite for those modal metaphysicians, including Lewis and Armstrong, who reject the possibility of nothing.[44] But there is worse to come.

The third and most important reason why the Truthmaker Principle is no part of fundamental metaphysics is that it is motivated by a wrong account of ontological grounding, wrong on two counts. Let us take the relation of ontological grounding to be a relation between propositions and propositions, with both arguments plural. Wherever one might speak of *entities* doing the grounding, we will instead say that the corresponding existence propositions provide the ground. (This parallels what was said in section 11.1 about the truthmaker relation.) When some propositions are fundamental, and sufficiently inclusive to ground all truths, call those propositions an *ontological base*; if there *is* an ontological base, say that *truth is ontologically grounded*. Now, although both Lewis and the truthmaker theorists agree that truth is ontologically grounded, the truthmaker theorist puts two conditions on ontological grounding that Lewis would reject. First, truthmaker theorists hold that for truth to be ontologically grounded there must be an ontological base consisting entirely of existence propositions. I will return to this below. Second, truthmaker theorists hold that *entailment* by propositions in the base is necessary for ontological grounding: truth is ontologically grounded only if every truth is *entailed* by some truths in the ontological base. Call this the *entailment constraint*. (Entailment may be strict implication, or may be something even stronger.) Lewis holds instead that only *supervenience*, not entailment, is necessary for ontological grounding: truth is ontologically grounded only if every truth *supervenes* on the ontological base. Call this the *supervenience constraint*. (Supervenience here is *global* supervenience applied to propositions.) If the ontological grounding of truth satisfies the entailment constraint, it also satisfies the supervenience constraint, but not vice versa. The supervenience constraint, then, is weaker than the entailment constraint, and demands less of the ontological base.

Truthmaker theorists, because they accept the entailment constraint, must deny that the atomic propositions form an ontological base. For example, negated atomic truths and general truths need not be entailed by the true atomic propositions. Thus, truthmaker theorists must add non-atomic propositions to the ontological base to ground negated atomic and general truths; Armstrong, for example, adds propositions involving totality states of affairs. Lewis, because he accepts only the supervenience constraint, disagrees: no propositions beyond the atomic propositions need be included in the ontological base. This is because all truths supervene on the atomic truths; two worlds that agree with respect to the truth value of all atomic propositions must agree with respect to the truth value of all negated atomic propositions, of all general propositions, indeed, of all (qualitative) propositions. Speaking picturesquely: once God has fixed the truth value of all the atomic propositions, the (qualitative) nature of the world is thereby fully determined.[45]

The other condition that truthmaker theorists wrongly require for ontological grounding is that there be an ontological base consisting entirely of existence propositions. This condition is easily met by states-of-affairs theorists. Each fundamental proposition in the ontological base can be replaced by a necessarily equivalent existence proposition asserting that the corresponding state of affairs exists. For example, the atomic proposition that *a* is *F* can be replaced by the existence proposition that *a's being F* exists; the totality proposition that *a* totals *F* can be replaced by *a's totaling F* exists; and so on. Lewis, however, feels no compulsion to whittle down the fundamental propositions in the

ontological base to existence propositions.[46] One aspect of being involves *what* entities there are; but another aspect of being involves *how* those entities are, and *how* they are arranged. The atomic propositions by themselves give full expression to both of these aspects of being. Thus, if truth supervenes on the atomic propositions, then we can say that *truth supervenes on being*:

> **(TSB)** For any proposition P and any worlds W and V, if the same entities exist in W and V, and those entities instantiate the same fundamental properties and relations in W and in V, then P is true in W iff P is true in V.[47]

For an anti-haecceitist, the conditional simplifies to: if W and V have the same pattern of co-instantiation of fundamental properties and relations, then P is true in W iff P is true in V. According to Lewis, it is (TSB), not (TM), that is a fundamental principle of metaphysics.[48] Nothing more than (TSB) is needed to ensure that truth be ontologically grounded.

A defense of (TSB) must show that, although weaker than the Truthmaker Principle, it is still strong enough to do some ontological heavy lifting. For truthmaker theorists, an important role of the Truthmaker Principle, is to "catch cheaters": philosophers who, by positing truths without truthmakers, fail to own up to the ontological cost of their theories. Consider, for example, the phenomenalist who holds that propositions about physical objects can be analyzed in terms of sense-data.[49] To analyze propositions about *unobserved* objects, such as that a ball in an otherwise empty room is red, the phenomenalist typically calls upon the sense-data that an observer *would* have had, had she been in a position to observe the ball. But, the truthmaker theorist asks, what are the truthmakers for these counterfactual truths? Not the unobserved ball (or states of affairs involving it) because these, according to the phenomenalist, do not exist. And not any actual sense-data (or states of affairs involving them) because these are all compatible with the ball not being red. The phenomenalist counterfactuals, it seems, are "brutely true" in violation of the Truthmaker Principle.

Lewis is in rough agreement with the truthmaker theorists as to which philosophical views are guilty of cheating – except, of course, that he doesn't think the thing theorist, whether or not a flexible counterpart theorist, counts among the cheaters. It is thus imperative for Lewis that (TSB) be strong enough to rule out views that posit brute counterfactuals (as well as brute dispositions, brute lawhood, and brute tensed properties). And so it is. For example, on any plausible version of phenomenalism, the phenomenalist counterfactuals don't supervene on being: two worlds can agree on the truth value of all atomic propositions, propositions about (actual) sense-data, and yet disagree on the truth value of counterfactuals as to what sense-data would have existed, had unmade observations been made. (This is especially obvious if one grants that there are possible worlds without any observers, and so without any (actual) sense-data.) Thus Lewis, no less than the truthmaker theorists, can reject phenomenalism because it violates the dictum that truth be ontologically grounded.

The philosopher who invokes brute counterfactuals has a ready response. "Every view is entitled to choose its own base of fundamental propositions. I choose to take some counterfactual propositions to be fundamental. Perhaps these brute counterfactuals can be understood to attribute dispositional properties to the entities that populate the world. But, if not, they can always be taken to attribute fundamental properties to the world as a whole. If it is part of my view that the brute counterfactuals are fundamental propositions belonging to the ontological base, then my view is compatible with (TSB)." And a similar speech can be made by the other supposed cheaters.

Lewis would protest: "a philosopher is *not* entitled to take any proposition to be fundamental; the fundamental properties and relations involved in the fundamental propositions are all *categorical*; their essential nature is entirely intrinsic, given by their *quiddities*, not by their causal or nomological roles." And what justifies this view of the fundamental properties and relations? "Humean

recombination principles would fail if the fundamental properties and relations were not categorical."[50] This would seem to take us in a circle, since Humean recombination principles are only plausible if one assumes that the fundamental properties and relations are categorical, that their natures are intrinsic. A better Lewisian response is that the whole package is justified holistically by the success of the Humean metaphysics, and especially by its *perspicuity*. Its metaphysical rivals are shrouded in mystery, engendering at best an illusion of understanding.

It is now plain that the way in which *Truth Supervenes on Being* "catches cheaters," and puts constraints on metaphysical theorizing, has little to do with the "supervenience" part, and everything to do with the "being" part, and the conception of fundamentality that informs it. Oft-heard complaints that supervenience by itself is not a dependence relation, or a relation of ontological priority, though true enough, are beside the point. It is supervenience *on being*, where being is characterized in terms of the pattern of instantiation of fundamental properties and relations, that provides an ontological ground for truth. And it is the conception of fundamental properties and relations as *categorical* that gives (TSB) its ontological punch.[51] Those seeking an informative characterization of this Humean conception of fundamental properties will be disappointed; Lewis offers little more than gestures and hints. But one could scarcely exaggerate its importance throughout Lewis's writings. His entire Humean metaphysics is incomprehensible without it.[52]

Notes

1 See especially Armstrong 1989, 1997, and 2004.

2 I critically discuss this aspect of truthmaking theory in Bricker forthcoming a.

3 Lewis 1986c. This was a reply to comments delivered by Peter Forrest and David Armstrong on Lewis 1986b.

4 Lewis 1992 and 1998, respectively.

5 See also Lewis 1994a and 1994b for statements and endorsement of the idea that truth supervenes on being. Lewis uses 'fundamental' and 'perfectly natural' interchangeably; see Lewis 2009 for the terminological pros and cons. In this essay I stick with 'fundamental'.

6 Here a_1, a_2, ... are (one or more) particulars and R is a (monadic or polyadic) universal. Since the universals are *sparse* on Armstrong's theory, so are the atomic propositions, and the atomic states of affairs. On the distinction between sparse and abundant conceptions of properties and propositions, see Lewis 1986a, 59–63.

7 This contrasts with Armstrong's account, according to which truthmaking is a primitive, cross-categorial relation (2004, 5). Armstrong argues against reducing the truthmaker relation to *entailment* (p. 6). But his argument depends upon taking entailment to be a more fine-grained relation between propositions than is strict implication, and so would not apply to Lewis's account.

8 For suggestions along these lines, see Restall 1996.

9 Restricting the scope of (TM) to contingent propositions, however, may not succeed in sidestepping the problem if there are distinct contingent propositions that are necessarily equivalent; any two such propositions could not differ with respect to their truthmakers. (On Lewis's own account of propositions, according to which propositions are classes of possible worlds, necessarily equivalent propositions are identical; but Lewis's professed neutrality prohibits him from presupposing this account.) The conjunction of any necessary proposition with a contingent proposition is contingent. Once we allow that there are distinct necessary propositions that differ with respect to their truthmakers, it is hard to see how the conjunctions of these necessary propositions with a contingent proposition would not also be distinct propositions that differ with respect to their truthmakers. So if there was a problem to begin with, simply restricting (TM) to contingent propositions won't solve it. The right way to sidestep the problem is this: first, form a quotient algebra of the Boolean algebra of propositions by identifying necessarily equivalent proposition; then, second, restrict the scope of (TM) to the resulting equivalence classes of contingent propositions, what might be called the

thoroughly contingent propositions. All quantification over propositions in what follows is tacitly restricted to the thoroughly contingent propositions.

10 As Lewis (2001a, 608) notes, this requires that none of the *T*s is itself a proper class, since proper classes are not members of proper classes. Lewis provides a way around the problem.

11 Unless, perhaps, one is a priority monist. See Schaffer 2010.

12 We need not concern ourselves, when we say '*a* is *F*', with whether '*a*' refers to the thick particular *a*+ or the thin particular *a*; either way, the meaning of the predicate 'is *F*' can be adjusted to make the truth conditions of the whole sentence come out right.

13 Here, however, I wonder what justifies Armstrong's identification of a conjunction of atomic states of affairs with a state of affairs having a conjunctive universal as a constituent.

14 See Bricker (2006, 256–62) for a more nuanced discussion of the argument that things cannot be truthmakers for general propositions. See Lewis (1986a, 86–90) for the (weak) Humean principle of recombination that underlies the argument.

15 Two notes: First, Armstrong says "aggregate" rather than "sum," but elsewhere identifies aggregates with (mereological) sums. Second, I will suppose that the second *relatum* of the totalling relation is always a universal; it is unclear how else to justify the existence of the corresponding totality state of affairs.

16 Armstrong gives a somewhat different, but equivalent, account of how totality states of affairs provide truthmakers for general propositions in Armstrong 2004, 74. In any case, I see a problem. If the *F*'s overlap in such a way that some *F* is mereologically composed of parts of other *F*'s, then it is possible for a sum of *F*'s to total *F* without being all the *F*'s, and the account fails. This problem could be solved by taking the first *relatum* of the totalling relation to be a class, rather than a sum; though this raises further issues about the ontological status of classes, and states of affairs involving classes.

17 In Lewis 2001a he explicitly endorses what he there calls a "deflationary theory of truth." And although his views changed somewhat between Lewis 2001a and Lewis 2001b (see the following note), I know of no reason to think that what is called "the redundancy theory" in Lewis 2001b isn't just the same theory renamed.

18 It is worth noting that Lewis's discussion of theories of truth in 2001b differs in substantial ways from his discussion in 2001a. (Although both papers were published in the same year, and neither refers to the other, I suppose that (2001b) contains his later, considered view. Lewis 2001a was circulated in draft form in May 1998, and presented at AAP in July 1998; Lewis 2001b was published in *Analysis*, which typically has a quick turnover.) In Lewis 2001a, the correspondence theory of truth is included among "the grand theories of truth"; and it is claimed that *all* the grand theories are compatible with the redundancy biconditionals, and so none the grand theories of truth are really about truth. In Lewis 2001b, as noted above, he argues that the pragmatic and coherence and other substantial theories of truth are incompatible with the redundancy biconditionals. It is because the correspondence theory (if there were such a thing) would presumably be taken in conjunction with the redundancy biconditionals that it, alone, turns out not really to be a theory of truth.

19 For an alternative, substantial discussion of Lewis's critique of the Truthmaker Principle with different points of emphasis than what follows, see MacBride 2005.

20 There is one sort of actualist, however – an actual world exceptionalist – who holds that in alien possible worlds, existential propositions may lack truthmakers.

21 Note that (DM) demands that there be no indiscernible possible worlds, since indiscernible worlds do not differ with respect to what exists in them. If one wants to avoid this consequence, one can restrict the Truthmaker Principle to what Lewis (oddly) calls *discerning* propositions, where a proposition is *discerning* iff it is never true at one but not the other of two indiscernible worlds: only discerning propositions have truthmakers. (Correlatively, the initial quantifier in (DM) is 'for any two *discernible* worlds *W* and *V*'.) See Lewis 2001a, 606–7. Note that for an anti-haecceitist a proposition is discerning if and only if it is qualitative. For Lewis's affirmation of anti-haecceitism, see Lewis 1986a, 220–35.

22 See Lewis 1991. Other potential counterexamples to uniqueness of composition – for example, involving compositional change over time – are also discussed and dismissed (pp. 78–9).

23 'Distinct', in this context, means non-overlapping, rather than non-identical; I trust that context successfully resolves this ambiguity throughout the essay. Lewis's statement of the principle of recombination is

rough, and in need of qualification. For example, distinct duplicates cannot "fail to coexist," as Lewis understands that phrase. For a detailed attempt to set all this right, see Bricker forthcoming b.

24 For a more detailed discussion of the relation between the Humean prohibition and composition, see Bricker forthcoming c.

25 Counterpart theory was introduced in Lewis 1968. It was expanded and modified in Lewis 1971, 1973, and 1986a.

26 Note that if *a* has multiple counterparts inhabiting *W*, some of which are *ø* and some of which are not *ø*, then *W* represents *de re* of *a* both that *a* is *ø* and that *a* is not-*ø*, although not, of course, that *a* is *ø-and-not-ø*.

27 See Lewis (1986a, 10–13) for some limitations on these definitions. It is controversial, to be sure, whether essence should be analyzed in terms of modality; see Fine 1994. But here I follow Lewis in presupposing it.

28 Lewis introduced multiple counterpart relations to solve the problem of contingent identity in Lewis 1971.

29 This deviates from what one would get if one slavishly applied the translation scheme from Lewis 1968 to (TM) (even putting to one side the introduction of multiple counterpart relations). As discussed in Lewis (1986a, 10–13), the translation scheme does not give the expected results when applied to sentences that are, or contain, existence propositions.

30 One might wonder whether a similar treatment could be applied to structural universals, thus taking them to be mereologically composed. Such a "reconsideration" of structural universals would undercut Lewis's reason for rejecting universals in Lewis 1986b. But if there are, or could be, basic laws involving structural universals, then such a reconsideration would not be compatible with either Lewis's or Armstrong's theory of laws.

31 A third, and I think better, reply is to hold that fundamental relations are necessarily symmetric; but I know of no reason to think Lewis would support it. Dorr (2004) provides some arguments for this view.

32 Material from this section was presented in a lecture entitled "The World: Facts or Things?" at NYU in February, 1999.

33 Lewis (2001b, 279) does say that the Truthmaker Principle "is equivalent" to an infinite bundle of biconditionals so as to emphasize the dispensability of the notion of truth; see §3 above. But elsewhere, he formulates the Truthmaker Principle as a single assertion by quantifying universally over propositions. The argumentation of Lewis (2001a) depends on it.

34 This requires two assumptions. First, that duplicate worlds are qualitatively indiscernible requires that worlds are externally isolated, that no part of any world is externally related to any part of any other world. (I defend this in Bricker 1996.) Second, that qualitatively indiscernible worlds are indiscernible *tout court* requires anti-haecceitism (see Lewis 1986a).

35 *Proof.* Suppose *a* is not *ø*, for intrinsic *ø*; and let *W* be a world where *a* exists, that is, where *a* has counterparts$_T$. By (11.1) and (11.3), all of *a*'s counterparts$_T$ in *W* are thick particulars and duplicates of *a*+. Since they are thick particulars, they have a property *ø* iff they *include ø*. Since they are duplicates of *a*+, they do not include *ø*. Therefore, no counterpart$_T$ of *a*+ has *ø*, and the proposition that *a is not ø* is true in *W*. Which goes to show: *a* is a truthmaker for *a is not ø*.

36 A relation is *intrinsic* if it is either *internal* or *external*, and thus supervenes on the intrinsic natures of its *relata*, taken together. See Lewis 1986a, 62. Lewis supposes that fundamental properties and relations are intrinsic. (See section 11.6.)

37 See Hazen 1979 and Lewis 1986a, 232–3.

38 Lewis (1986a, 232–3) calls the counterpart pair $<a',b'>$ a "joint possibility" for $<a,b>$. So (11.4) could be rephrased: if *W* contains individual possibilities for *a* and for *b*, then it contains a joint possibility for $<a,b>$; and joint possibilities preserve intrinsic relations between *a* and *b*.

39 See Hazen 1979, 333–4. For the general case, Hazen quantifies over the "representative functions." The modification in the semantics is needed in any case to solve the problem of "essential relations."

40 *Proof.* Suppose *aRb*, for intrinsic *R*. Consider any world *W* where *a* + *b* has a counterpart$_T$. By (ME), *a* and *b* have counterparts$_T$ in *W*. By (11.4), $<a,b>$ has a counterpart$_T$ pair in *W*, $<a',b'>$ where $<a,b>$ and $<a',b'>$ are intrinsic isomorphs. Therefore, since *aRb* and *R* is intrinsic, *a′Rb′*. Thus, *aRb* is true in *W*, as was to be shown.

41 There are various ways in which a flexible counterpart theorist can make precise the notion that "one entity can count as two" in intensional contexts. And, of course, truthmaking contexts, on modal accounts of truthmaking, are intensional contexts.

42 Would it help to take states of affairs to be sets – sequences of universals and (thin) particulars – instead of mereological sums? No, on any states-of-affairs theory, the (concrete) world is composed of states of affairs; thus, if states of affairs are sets, the (concrete) world itself, implausibly, would have to be a set, or composed of sets.

43 See Bricker (2001, 47–9) on how a realist about possible worlds can accommodate the possibility of nothing.

44 Lewis nonetheless transforms it into a meta-metaphysical objection. The Truthmaker Principle, he writes, provides "a swift reason why there must be something, and not rather nothing … Altogether too swift, say I" (Lewis 2001a, 611).

45 The modifier 'qualitative' is needed because indiscernible worlds (if any) agree with respect to the truth value of all atomic propositions. Note that the atomic propositions must include all *possible* atomic propositions, including those (if any) involving alien universals.

46 No compulsion. But a thing theorist can do this without cost by interpreting existence propositions using the duplication relation as the counterpart relation: a thing exists in all and only those worlds that have a duplicate of that thing as a part. This leads to a stronger version of Truth Supervenes on Being that I call the Subject Matter Principle: every proposition has a subject matter, entities such that the truth or falsity of the proposition is determined by whether or not those entities exist. See Bricker 2006, 270–7.

47 This is essentially the formulation in Lewis 2001a, 612. See also Bigelow 1988, 38. (TSB) is sometimes called "Truthmaker" in the literature. Although (TSB) can be understood as characterizing a weaker truthmaker relation, I think it best to only speak of truthmaking if the truthmaker *necessitates*, or *entails*, the truth.

48 Must one be a modal realist to take (TSB) to be a fundamental principle of metaphysics? Lewis thought not; he took the ontological dispute between (TM) and (TSB) to cut across disputes in the metaphysics of modality. But, certainly, some actualist reformulations of (TSB) – by linguistic ersatzists, for example – are not plausible candidates for fundamental metaphysical principles; and some actualists lack the modal means to provide any formulation of (TSB).

49 Armstrong uses this example to introduce truthmaking in Armstrong (2004).

50 See Lewis 1986a, 162–3. Sider (2012, 155) suggests that replacing (TSB) with "the correct fundamental ideology is that of predicate logic" would have "essentially the same upshots regarding cheaters." But I don't see how that would catch "cheaters" who gladly trade their fundamental propositional operators for fundamental properties of worlds.

51 Lewis's most extensive discussion of the nature of fundamental properties is in Lewis 2009. Lewis does not much use the word 'categorical'; perhaps a better label would be 'Humean', since, for Lewis, only properties that satisfy Humean recombination principles are candidates for being fundamental.

52 Thanks to Jonathan Schaffer for helpful comments.

References

Armstrong, D.M. (1989). *A combinatorial theory of possibility*. Cambridge: Cambridge University Press.

Armstrong, D.M. (1997). *A world of states of affairs*. Cambridge: Cambridge University Press.

Armstrong, D.M. (2004). *Truth and truthmakers*. Cambridge: Cambridge University Press.

Bigelow, J. (1988). Real possibilities. *Philosophical Studies*, 53, 37–64.

Bricker, P. (1996). Isolation and unification: the realist analysis of possible worlds. *Philosophical Studies*, 84, 225–38.

Bricker, P. (2001). Island universes and the analysis of modality. In G. Preyer & F. Siebelt (Eds.), *Reality and Humean supervenience: essays on the philosophy of David Lewis*. Lanham, MD: Rowman & Littlefield.

Bricker, P. (2006). The relation between general and particular: entailment vs. supervenience. In D. Zimmerman (Ed.), *Oxford Studies in Metaphysics*, Vol. 2, pp. 251–88. Oxford: Oxford University Press.

Bricker, P. (forthcoming a). Ontological commitment. *Stanford encyclopedia of philosophy*.
Bricker, P. (forthcoming b). Principles of plenitude. In *Modal matters: essays in metaphysics*. Oxford: Oxford University Press.
Bricker, P. (forthcoming c). Composition as a *kind* of identity. *Inquiry*.
Dorr, C. (2004). Non-symmetric relations. In D. Zimmermann (Ed.), *Oxford Studies in Metaphysics*, Vol. 1, pp. 155–93. Oxford: Oxford University Press.
Fine, K. (1994). Essence and modality. *Philosophical Perspectives*, 8, 1–16.
Gibbard, A. (1975). Contingent identity. *Journal of Philosophical Logic*, 4, 187–222.
Hazen, A. (1979). Counterpart-theoretic semantics for modal logic. *Journal of Philosophy*, 76, 319–38.
Lewis, D. (1968). Counterpart theory and quantified modal logic. *Journal of Philosophy*, 65, 113–26.
Lewis, D. (1971). Counterparts of persons and their bodies. *Journal of Philosophy*, 68, 203–11.
Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.
Lewis, D. (1986a). *On the plurality of worlds*. Oxford: Blackwell.
Lewis, D. (1986b). Against structural universals. *Australasian Journal of Philosophy*, 64, 25–46. Reprinted in Lewis (1999).
Lewis, D. (1986c). A comment on Armstrong and Forrest. *Australasian Journal of Philosophy*, 64, 92–3. Reprinted in Lewis (1999).
Lewis, D. (1991). *Parts of classes*. Oxford: Blackwell.
Lewis, D. (1992). Critical notice of Armstrong, *A combinatorial theory of possibility*. *Australasian Journal of Philosophy*, 70, 211–24. Reprinted as "Armstrong on combinatorial possibility" in Lewis (1999).
Lewis, D. (1994a). Reduction of mind. In S. Guttenplan (Ed.), *A companion to philosophy of mind*. Oxford: Blackwell. Reprinted in Lewis (1999).
Lewis, D. (1994b). Humean supervenience debugged. *Mind*, 103, 473–90. Reprinted in Lewis (1999).
Lewis, D. (1998). Critical notice of Armstrong, *A world of states of affairs*. *Times Literary Supplement* 4950, 30. Reprinted as "A world of truthmakers" in Lewis (1999).
Lewis, D. (1999). *Papers in metaphysics and epistemology*. Cambridge: Cambridge University Press.
Lewis, D. (2001a). Truthmaking and difference-making. *Noûs*, 35, 602–15.
Lewis, D. (2001b). Forget about the "correspondence theory of truth". *Analysis*, 61, 275–80.
Lewis, D. (2003). Things *qua* truthmakers. In H. Lillehammer & G. Rodriguez-Pereyra (Eds.), *Real metaphysics: essays in honor of D.H. Mellor*, pp. 25–38. London: Routledge.
Lewis, D. (2009). Ramseyan humility. In D. Braddon-Mitchell & R. Nold (Eds.), *Conceptual analysis and philosophical naturalism*, pp. 203–22. Cambridge, MA: MIT Press.
Lewis, D. & Rosen, G. (2003). Postscript to "Things *qua* truthmakers". In H. Lillehammer & G. Rodriguez-Pereyra (Eds.), *Real metaphysics: essays in honor of D.H. Mellor*, pp. 39–42. London: Routledge.
MacBride, F. (2005). Lewis's animadversions on the truthmaker principle. In H. Beebee & J. Dodd (Eds.), *Truthmakers: the contemporary debate*,(pp. 117–40. Oxford: Oxford University Press.
Restall, G. (1996). Truthmakers, entailment and necessity. *Australasian Journal of Philosophy*, 74, 331–40.
Schaffer, J. (2010). The least discerning and most promiscuous truthmaker. *Philosophical Quarterly*, 60, 307–24.
Sider, T. (2012). *Writing the book of the world*. Oxford: Oxford University Press.

12

How to Be Humean

JENANN ISMAEL

> *There are no ideas, which occur in metaphysics, more obscure and uncertain, than those of power, force, energy or necessary connexion.*
>
> (Hume 1902, 61–2)

The Doctrine and the Program

David Lewis famously said in the introduction to his second volume of *Philosophical Papers* that he saw a lot of his career in retrospect as being devoted to the defence of Humean supervenience (HS), the metaphysical doctrine that "all there is to the world is a vast mosaic of local matters of particular fact, just one little thing and then another."[1]

The research program as he conceived it was to provide truth conditions for all contingent truths in terms of what he came to call the Humean mosaic.

I have become increasingly confused over the years about what Humean analyses are supposed to achieve. I will argue that if they are supposed to provide content-preserving reductions, they fail for one reason. If, on the other hand, they are supposed to tell us what it is in the realm of Being, according to the Humean, that our beliefs about various things – laws, chances, the value of a dollar bill, or the beauty of a sunset – refer to, they fail for different reasons. I will suggest a significant shift in how the Humean research program is conceived.

In the first section, I argue that Humean analyses don't provide content-preserving reductions. In the second section, I argue that Humean analyses don't provide non-trivial accounts of the reference. In the remaining sections, I introduce a distinction between structure in the realm of Being and structure in our representations of Being and argue that there are good reasons not to expect content-preserving reductions of the modal to the non-modal at the level of content, or useful mappings of content-level structures into structures at the level of Being. I spend the rest of the paper groping towards a reconception of the relationship between content-level structure and structure at the level of Being. I will argue that the debate between the Humean and non-Humean should be

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

reconceived as a debate about whether non-Humean facts play a substantive role in the story of how beliefs about laws and chances are formed and used.

These are issues that challenge assumptions about the way that content-level structure relates to structures at the level of Being built into practices in analytic metaphysics that Lewis himself had a substantial hand in developing. His work in every area of analytic metaphysics has stood as paradigmatic for a generation for good reason. I think of the shift that I am suggesting in how the program is conceived as something that will advance the program.

12.1 Humean Supervenience and the Failure of Content-Preserving Reduction

There are reasons for worrying about the claim that all there is in the world is a vast mosaic of local matters of particular fact stemming from the apparently non-local correlations connected with quantum phenomena. I will not be addressing those, although they raise important issues.[2] Aside from problems with quantum non-locality, the biggest challenge for HS comes from modalized concepts that are central to science: laws and chances.[3] Lewis devoted a substantial part of his career to Humean analyses of these concepts and his analyses are widely regarded as triumphs of philosophical analysis. Humean supervenience is, as Lewis remarks, named after the great denier of necessary connections in Nature. The difference between Humeans and non-Humeans is that the non-Humean sees patterns in the Humean mosaic as *distinct from*, and typically *explained by*, facts about laws and chances. The Humean, by contrast, sees patterns in the Humean mosaic as *constituting such facts*.

There is a massive literature on details of Lewis's analyses. But the primary motivating factors for the Humean view are that it is metaphysically modest and fits very well with scientific practice. If beliefs about laws and chances and such are veiled beliefs about patterns in the Humean mosaic, that explains very nicely both how we arrive at defeasible beliefs about them and the role they play guiding expectation. It gives us a nice account of confirmation and explains scientific practices and proves remarkably adept at navigating its way through the minefield of objections that have been leveled against it. It is not an accident that physicists and philosophers of science tend to be Humeans.

Lewis's analyses build on a long tradition that includes regularity accounts of law and frequency accounts of chance. Regularity accounts of law hold out the promise of reducing claims about laws to claims about regularities. In its simplest form, a law is a true, universally quantified proposition that holds without exception. Frequentist accounts of chance hold out the promise of reducing claims about chance to claims about the frequencies, or relative frequencies of events (really, event *types*) in a certain kind of reference class. Both types of account, in their simplest form, suffer from obvious counterexamples, and have undergone refinements. The Humean wants to refine them without invoking modality in an ineliminable role. Since he is looking to give truth conditions that invoke only patterns in the Humean mosaic it wouldn't do the Humean any good, for example, to respond to counterexamples to regularity accounts of law by saying laws are not only those regularities that hold in fact, but also those regularities that hold with physical *necessity*. Or to respond to counterexamples to frequency accounts of probability by identifying chances not just with actual frequencies, but with long-run *hypothetical* frequencies. Modal facts are supposed to be part of the output of a Humean analysis, not part of the input.

12.1.1 Best Systems

Lewis, following leads from Ramsey and Mill, saw the need to move from an attempt to give conditions that individual regularities have to satisfy to qualify as laws to holistic reduction of the *systems* of laws associated with a physical theory. The early view ran as follows.

> Take all deductive systems whose theorems are true. Some are simpler, better systematized than others. Some are stronger, more informative than others. These virtues compete: an uninformative system can be very simple; an unsystematized compendium of miscellaneous information can be very informative. The best system is the one that strikes as good a balance as truth will allow between simplicity and strength. . . . A regularity is a law iff it is a theorem of the best system.[4]

Later, he extended the account to accommodate chance. In the new view, a claim about the laws and chances that obtain at our world formed a theoretical package that could be reduced only all at once to a claim about patterns in the Humean mosaic.

What the Best System account (BSA) gets right is the holistic epistemology of beliefs about laws and chances. What Lewis describes is, more or less, an accurate schematic description of the standards that govern choice of a fundamental theory in physics. The theorist doesn't start with a full description of the Humean mosaic, but he uses information about the Humean mosaic and generates hypotheses about laws and chances in a way that is governed by norms of simplicity and strength of the kind that Lewis was pointing to. It is a great virtue of Lewis's account that, instead of consulting his intuitions about what laws and chances are, he looked to science and gave a recognizable description of how claims about laws and chances are generated from information about actual facts.

Where the BSA goes wrong is that it fails to capture the truth conditions right for claims about laws and chances for something like the same reasons that regularity accounts fail to capture the truth conditions right for claims about laws and frequency accounts fail to capture the truth conditions for claims about probability. The problem is a straightforward generalization of a characteristic problem faced by the more simple-minded regularity accounts of law and frequency accounts of chance. The problem with those accounts is that they seem to fail the test for identity. If facts about laws are constituted by facts about regularities, there ought not to be possible worlds in which the facts about laws come apart from the facts about regularities.[5] But no one that doesn't recognize the possibility of regularities that are not laws is a competent user of the notion of law. And it doesn't matter how fancy you get in characterizing the type of regularity in question. And if facts about chances are constituted by facts about frequencies, there ought not to be possible worlds in which the facts about chances come apart from the facts about frequencies. But no one that doesn't recognize the possibility of divergence between chance and frequency is a competent user of the notion. In this case, the possibility of chances that diverge arbitrarily far from the frequencies is explicitly codified in the axioms of probability. Bernoulli's theorem says that it doesn't matter how large your sample is, the possibility that the chance of an event will diverge from its frequency in the sample remains. The logic of law beliefs (by which I mean their inferential relations to beliefs about local matters of particular fact) explicitly recognizes the possibility of worlds in which there are regularities (even exceptionless regularities) that are not laws and the chances diverge from the frequencies. And any competent user of those concepts has to recognize those possibilities.

The problem for the BSA is quite similar. There is a modal gap between the facts about laws-and-chances and the categorical facts. It is part of the logic of those concepts, embodied in the norms that govern inferences with them. In physical contexts, to say that x is possible according to theory T just is to say that there is model of T in which x is the case. And we can show that there are models of our law + chance packages in which the categorical facts are very different than they actually are, and there are worlds in which the categorical facts are as they actually are, but the laws and chances are different. Let C be the categorical facts at a world W, and T the best systematization of C. There are worlds at which C but not T (C is a model of other theories), and worlds at which T but not C (there are models of T in which not C). It is sometimes remarked that in T-worlds that are not C, the laws are different as well. It is unclear that this is so. Non-C T-worlds are still *T-worlds*, even if T is *not* the best systematization of the facts at those worlds. The situation is exactly the same as in

the link between laws and regularities, or chance and frequencies. At the level of belief, categorical facts are treated as *evidence* for beliefs about chance, but there is no reduction. Just as the truth conditional content of claims about probability explicitly recognizes the possibility of frequencies that diverge as far as you please from the probabilities, the truth conditional content of T explicitly recognizes the possibility of T-worlds in which T is not the best systematization. To say that L is a law is to add something to the claim that it is an exceptionless regularity. To say that an event e has a high chance of occurrence is to say both more and less than that it actually occurs. And likewise, to say that T is the correct theory is to say both more and less than C. It is to say less than C because there are worlds in which T and not C, and it is to say more, because there are worlds in which C and not T.

This sort of argument is not new. Maudlin gives a version of it in application to the laws:

> Let us suppose (and how can one deny it) that every model of a set of laws is a possible way for *a world governed by those laws* to be. Then we can ask: can two different sets of laws have models with the same physical state? Indeed they can . . . The situation is even worse for probabilistic laws. Consider a law that assigns a probability to any given event, say the decay of a radioactive atom. The models of such a law will include worlds where every decay event assigns a different probability to the event . . . again since different laws share the same models, either the laws cannot supervene on the matters of particular fact or else some models of the laws cannot be regarded as physical possibilities relative to those laws.[6]

Why is this not a knockdown argument? Lewis never to my knowledge addressed this form of argument. Loewer holds that arguments that point to putatively possible worlds in which beliefs about law + chance packages and beliefs about their putative Humean truthmakers come apart does little more than express "intuitions of non-supervenience." "The only 'evidence' that the anti-Humean can point to that would, without begging the question, count in favour of the existence of [non-Humean truthmakers for law statements] is our intuitions of nonsupervenience."[7] And he thinks that these intuitions have little probative value:

> There is a long tradition in philosophy of evoking intuitions that are associated with concepts in order to discover the nature of the concepts' reference. This method seems more appropriate for some concepts than for others. But . . . when [the] subject matter is scientific and when the intuitions concern modality, the argument is very weak and easily defeated by alternative explanations of why we have the intuitions we do.[8]

He argues that, even if they do not capture all of our pre-theoretic beliefs about laws, Lewis-style Humean replacements for laws serve all of the important scientific functions of laws, and we can reject the anti-Humean intuitions without loss and with the reward of a sensible ontology.

But that is not what is going on here. The problem with the BSA is that it tries to understand a physical theory as providing a summary of information about the manifold of actual events, but we don't get the inferential calculus in which beliefs about laws and chances are related to beliefs about matters of actual fact right if we identify them with beliefs about patterns in the Humean mosaic. Knowing how to reason with claims about laws and chances demands recognition that even if it is a good *epistemic* inference from a pattern in the manifold of actual fact to a belief about laws or chances the logic of the inference undermines any equivalence by explicitly allowing for the possibility of worlds in which that inference fails. This is so even *more* explicitly for reasoning in the context of a physical theory. In such a context we don't have to rely on intuitions about what is possible. Physical possibility is represented by the theory's models and we can point to models that falsify the supervenience claim. The content of claims about laws and chances clearly distinguishes them from any set of claims about actual facts.

To think otherwise is to make the same mistake that the behaviorists make with respect to mental facts: that is, to go from the innocent and undeniable observation that behavior forms the basis for beliefs about mental states to the view that such *represent* complicated facts about behavior. Understanding what attributions of mental states *mean*, in a sense that can be understood purely in terms of their inferential connections to other beliefs, demands recognition that the inference from behavior to mentality is defeasible in both directions. The doxastic outputs of the BSA cannot be understood in purely categorical terms any more than the doxastic outputs of a theory of mind can be understood in purely behavioral terms. The BSA undoubtedly provides acceptance conditions, but the connection it establishes between beliefs about patterns in the Humean mosaic and beliefs about laws and chances falls short of equivalence. It is – as we say – an *ampliative* inference.

This shows that if the BSA may provide a goods account of how beliefs about laws and chances are generated from beliefs about the pattern of actual fact, it doesn't provide a content-preserving reduction of modal claims to non-modal ones. Loewer wants to say that maybe, in our everyday reasoning, we treat beliefs about laws and chances *as though* they were separable from beliefs about the pattern of actual fact, and maybe *as though* they explained those patterns, but really any claim about the laws and chance is just a redescription, a summary, a compact way of saying something very general about the pattern of actual fact. This is, of course, the same kind of thing that the behaviorists say about claims about the mental. There are two things to say. First, it is elimination, not an analysis. More importantly, if we add the inference that the Humean wants us to add to our theory – that is, the inference from a description of the Humean mosaic to (L + C) – we close a modal gap that the theory itself recognizes, we eliminate models that the theory itself recognizes as physical possibilities. That means that adding that inference is not a conservative extension, and *that* is enough to show that the BSA can't be a content-preserving reduction. It doesn't get the content of law and chance claims right even within the inferential calculus provided by a theory.

12.1.2 The Function of Beliefs about Law and Chance

There is a welcome emphasis on the pragmatic motivation for theorizing in Lewis's discussions of the BSA that is even more explicit in the discussions of contemporary Humeans. David Albert, for example, makes it clear that the BSA is a description of nature that answers to the pragmatic needs of agents with our limitations. He introduces the BSA as a response that God would make to the following kind of demand:

> Imagine that you have an audience with God who provides you with as much information about the particular facts of the world as you could possibly want to have. One way to provide with information is to recite long lists of particular facts concerning which properties are instantiated at which spatiotemporal locations. Yet, as God begins to recite particular matters of fact, it begins to look as if all this is likely to drag on for a while. And you explain to God that you're actually a bit pressed for time, that this is not all you have to do today, that you are not going to be in a position to hear out the whole story. And you ask if maybe there's something meaty and pithy and helpful and informative and short that He might be able to tell you about the world which (you understand) would not amount to everything, or nearly everything, but would nonetheless still somehow amount to a lot. Something that will serve you well, or reasonably well, or as well as possible, in making your way about in the world.[9]

There is a focus on summarizing in these remarks, but that's not quite enough. Scientific theories represent in terms that are designed to facilitate more specific practical and epistemic tasks and they are formed under epistemic constraints. Criteria for choice of best system (the criteria by which best-ness is judged) reflect this function. These will give content to what is meant by "meaty and pithy and helpful and informative and short." So, for example, the information that goes *into* the

construction of the best system has to be discoverable by limited beings, and the packaging it affects gives it a shape that answers to human needs. To be informative they have to be stated in a language from which we can extract information about the things we are interested in. To be helpful, they have to help us solve the kinds of problems we frequently encounter. Meatiness, pithiness, and shortness can be measured by strength and compactness. They also have to be accessible to the likes of us under the kinds of epistemic conditions we face. If the only way of forming (not necessarily infallible, but reasonable, defeasible) opinions about best systems were to have a private audience with God, they wouldn't provide practicable solutions to real problems.[10] The point is that the criteria by which best systems are judged need to be filled out in terms that are relative to human abilities and ends.[11] A better understanding of the kinds of cognitive tasks that BSAs subserve gives more specific content.

Cognition is in every way shaped and designed to help us overcome the informational bottleneck created by the low bandwidth of perception with an eye to facilitating effective action. Unlike precognitive systems that respond to information contained in sensory states, we process sensory information through an evolving model of the world that serves as a repository for information acquired from multiple sources over a whole history of experience. These models contain all kinds of structures poised to facilitate practical and empirical inferences, without any clear distinction between those that perform a purely descriptive function and those that have some more practical role. Science is an extension of cognition, with the same epistemic and practical interests. In these terms, we can see more clearly why the structures that play those roles can't be given content-preserving Humean reductions. Functionally, information about laws and chances are partially prepared inferences. They provide recipes for generating a good guess or strategies for making a good choice. To serve that function, they have to have implications not only for what *will* happen, but what would happen under conditions that may or may not be realized. Their modal implications outrun any summary of categorical facts and that play a crucial role in the kinds of cognitive tasks for which they are employed. The modal gap – that is, the ineliminable looseness of fit we saw between chance and frequency and also between law and regularity – has to do with the role of these shortcuts for solving practical and epistemic problems with the kind of loose fit to the facts that they need to play that role.

Consider the role that chances play guiding the formation of expectation in the absence of specific information about the outcome. In Lewis's framework, chances are single-case probabilities derived from history-to-chance conditionals which give the general probability of a particular event conditional on the history (or in a relativistic setting, the back light cone) up to the moment at which the event is slated to occur. Since chances are supposed to guide opinion in the absence of specific information from the future, these general probabilities can't be defined only for the actual future. They have to cover an open-ended class of potential futures, specifically all of those potential futures that are epistemically possible for the agent. Or consider the role that laws play in practical reasoning. An agent faced with a choice about how to act uses the laws to predict the downstream consequences of potential actions and her assessment of those consequences guides their choice. To play that role, they need to support conditional expectations about what *would* occur under alternative choices. And that means that no reduction that eliminates the modal content of law and chance beliefs could play the role they play in choice.[12]

The BSA does a very good job of capturing how beliefs about local matters of particular fact are transformed into beliefs about laws and chances. But the functional role of beliefs about law and chance demands that the content of those beliefs outrun any beliefs about what actually happens. This means that as an account of acceptance conditions, the BSA succeeds. But Lewis is absolutely clear that he intends it as a constitutive account of what laws and chances are. In his words: "I insist that I am not talking about how evidence determines what's reasonable to believe about laws and chances. Rather, I'm talking about how nature – the Humean arrangement of qualities – determines

what's true about the laws and chances."[13] He is equally clear that an analysis of what chance is has to answer to its role in cognition. As he puts it: "A feature of Reality deserves the name of chance to the extent that it occupies the definitive role of chance; and occupying the role means obeying PP." PP, of course, is the Principal Principle, which Lewis intended to capture the epistemic role of chance. It says that chances have the role of guiding belief about future events in the absence of specific information about their outcome.[14] If it can be shown that beliefs about laws and chances are not interchangeable in cognition with beliefs about patterns on the Humean mosaic, by his lights that would be a reason to reject the analysis. He spends a lot of time wringing his hands about whether the possibility of undermining futures (epistemically possible futures that are incompatible with the present chances) disqualify the distributed patterns that serve as truthmakers for best systems from playing the cognitive role of chance. The worry there is that the facts about chance are "inadmissible," which is to say that they defeat the application of PP (briefly, chances guide credence in the absence of inadmissible information, but information about chances itself looks like it turns out to be inadmissible). Lewis thinks that a solution to the undermining problem was found, but I'm suggesting that he should have seen it as a symptom of the more general problem that that emerges very clearly when we view the role of beliefs about probabilities and law in the cognitive life of the agent. *In the epistemic and practical context in which beliefs about laws and chances are deployed, they are not interchangeable with any set of beliefs about what actually happens.* Law and chance beliefs are not interchangeable in inferences with beliefs about categorical facts. They have modal implications that outrun those of any set of categorical facts, and those modal implications play an indispensible role in practical and epistemic reasoning. That is enough to show that no content-preserving Humean reduction of such claims is possible.

The BSA is a recipe for generating beliefs about what will or would happen from beliefs about what has happened. It takes the partial knowledge we have about the past and generates inductive hypotheses in the form of best systems, or chance + law packages, or scientific theories that tell us not what *does* happen, but what *would* happen under a range of epistemically and practically possible conditions. To eliminate the modal content of these packages is to eliminate a part of the inductive content that plays a crucial role in epistemic and practical reasoning. No Humean replacement for beliefs about law and chance that recasts them as summaries of information about the categorical facts can play that role.[15] If the fight between the non-Humean and the Humean is conducted on this battleground – that is, as a fight about the *content* of claims about law and chance – the non-Humean wins.

12.2 A Different Conception of Humean Reduction: Identifying Truthmakers

Suppose we acknowledge that Humean analyses don't provide content-preserving reductions of law and chance claims. There is a somewhat different conception of what Humean reductions are supposed to accomplish that is closer to the way that Lewis seems to have conceived it that I will now consider. When Lewis raised the question about what chance is, he raised it as a question not about the content of beliefs about chance, but a question about semantics, that is, a question about what features of the world those beliefs *refer to*, or what facts about the world make beliefs about chance *true*. Lewis holds as a general metaphysical doctrine that every truth must be made true by some feature of Being, and so, if HS is the right account of Being, every truth must be somehow *made true* by the local arrangement of qualities.[16] The Principal Principle was introduced as an implicit definition that identified chance by its connection to belief and then he set about looking for something in the Humean account of Being that could act as truthmaker for beliefs about chance. The Principal Principle says that believers without crystal balls should adjust their credences to the chances where

they are known, and so this means that, as he says: "Whatever *makes it true* that the chance of decay [of a radioactive atom within a given interval] is 50% must also, if known, make it rational to believe to degree 50% that decay will occur."[17] The idea is that we can think of any class of beliefs as picking out some aspect of what there is (something, we might say, in the realm of Being) in a manner that is implicitly defined by their role in reasoning, and it is left to the metaphysician to give an explicit characterization of what those beliefs pick out. The question for the metaphysician interested in chance is "What could we assign as truthmakers to beliefs about chance from the perspective of Being that would make sense of their cognitive role?"[18] And the Humean answer has to be some pattern in the Humean mosaic.

But if that is the question Lewis set out to answer, he didn't end up answering that question in an informative way. He effectively gave up on an interesting semantic story in terms of non-trivial correspondence to specifiable features of the Humean mosaic when he moved from individual reductions of chance and law to reduction of the content of chance + law packages. The holistic restructuring effected by the BSA undermined the hope of identifying specific features of the Humean mosaic that could be assigned as reference to beliefs about chance. What he ended up saying effectively was that nothing less than the whole pattern of categorical fact serves as truthmaker for the BSA as a whole and each of its theorems.[19] And that means that if beliefs about chance refer to *anything* in a Humean world, they refer to *everything*. If we ask "What makes it true that the chance of radium decaying at a particular moment in history is 50 percent?" the answer is the fact that it is a theorem of the best systematization of the whole unfolding history of the world that the chance of decay is 50 percent. Literally. Period. If we ask "What makes it true that the chance that a spin-half particle entering a measuring device at another moment, in a different place will show an up result is 50 percent?" the answer is the same. If we ask "What makes it true that the fundamental laws prohibit acceleration of massive particles past the light barrier, or that it is possible for a human being to swim the English Channel?" the answer is the same. And so it goes for every theorem of the BSA.

What Lewis really ended up doing is giving us a schematic account – in the BSA – for forming modal beliefs. The BSA was a good schematic account of how best systems package information for useful deployment in practical and epistemic problems. And the pragmatic criteria that govern that repackaging were immediately recognizable to many of those who work in the philosophy of science as the right account of theory choice in science. What Lewis *didn't* provide was either a content-preserving reduction of modal to non-modal belief or a non-trivial account of reference.

12.3 Digging Deeper

Let's make a distinction between what I call *content-level structure* and structure in the realm of Being. Content-level structures are structures defined over the contents of representations, whether these are theories, maps, mental models, or beliefs. Structure in the realm of Being is structure that is intrinsic to the mind-independent fabric of Being. There's a very confusing ambiguity in the notion of truth conditions that runs through philosophy and that doesn't distinguish content-level relations from relations that bridge the gap between content-level phenomena and Being. Sometimes truth conditions are construed as part of formal semantics. In that capacity, truth conditional equivalence preserves inferential role. Sometimes, however, truth conditions are thought of as assigning extensions from one's account of Being.[20] The arguments that I gave in section 12.1 against construing Humean analyses as content-preserving reductions were really arguments against the interchangeability of claims about laws and chances with claims about patterns in the Humean mosaic at the level of content. At the level of content, there is a separability canonized in the inferential calculus that relates beliefs about laws and chances to beliefs about categorical facts. When we thought about

the function of beliefs about laws and chance, we saw why that was so. Laws and chances have an inductive modal content that plays a role guiding expectation and decision, they have implications not only concerning what *does* happen, but also what *would* happen in epistemically and practically possible situations. We then considered a different view of what Humean analyses were supposed to achieve, which conceived of them as bridging the space between content-level structures and Being. In this capacity, Humean analyses were conceived as assigning truthmakers to law and chance beliefs drawn from the Humean mosaic. And we saw that Lewis failed to find informative correspondences. He ended up saying that every truth is made true by the whole pattern of fact.

I want to look a little more deeply now into why Lewis failed to find useful correspondences. Here I want to take a side-on look at the relationship of content-level phenomena and Being, when the contents in question are those of the structured representations that situated agents use to steer by (whether we are talking about beliefs, or the internal maps that we form to steer by, or the structured representations embodied in a scientific theory) and Being is characterized, as Aristotle would have said, qua Being, which is to say, a thoroughly deperspectivalized, intrinsic description of reality often referred to nowadays, somewhat misleadingly, as the God's Eye view.[21] Lewis offered the Humean Mosaic as an account of what there is in this sense, and he thought that if he was right, for every structure at the level of content, there has to be some corresponding structure at the level of Being that that structure reflects.[22] There are localized correspondences between particular localized objects and occurrences at the level of Being and structures on our models. But there won't be these localized correspondences between localized structures at the level of Being and laws and chances. Each law and chance belief is going to be made true by the way things generally hang together. The extension is distributed, but the truthmaker relation is going to be the same.

There are two ways of understanding the idea that every true belief, or every structure on our scientific models of the world, reflects some structure at the level of Being. One is harmless, but empty. The other is substantive, and intuitive, and it has played an important role guiding imagination about representation, but wrong. The harmless way is the one that has become known as the deflationary interpretation. I'll say little about it here, except as an alternative that avoids the pitfalls of the substantive and intuitive interpretation. The substantive and intuitive interpretation I want to make explicit because seeing what is wrong with it is the best way to get the right view. The substantive and intuitive interpretation of the "every truth has a truthmaker thesis" holds that every structure at the level of content has the job of *standing for* some feature of the way things *are*, that is, as Bernard Williams used to put it, "there anyway." This is an idea of truthmaking built on the paradigm of what Huw Price calls "matching games."

> Imagine a child's puzzle book, designed like this. On the right side of each page there's a picture of a complex scene, on the left side a column of peel-off stickers. For each sticker – the Opera House, the Harbour Bridge, the koala, and so on – the child needs to find the corresponding object in the picture. The game is successfully completed when every sticker has been placed in its correct location.
>
> Now think of the right-hand side as the world, and the column of stickers as the set of statements we take to be true of the world. For each statement, it seems natural to ask what makes it true – what fact in the world has precisely the "shape" required to do the job. Matching true statements to the world seems a lot like matching stickers to the picture . . .[23]

The paradigm of representation built on the model of the matching game fails to be a fully general account of the relations between content-level structure and Being. There are many different reasons that we might fail to find correspondences: emergence (some high-level structures will be the product of complex interactions at lower levels and resist mapping onto any structure in a static representation of Being), holism (some will carry partial information about the whole pattern, but

not full information about any part), implicit relationality (some will be implicitly relativized to the situation of a participant in Nature and appear only in a relational form). The reason that it fails when we look at structures like laws and chances is a little more general. It has to do with the job these structures play.

Laws and chances act in our cognitive lives as partially prepared solutions to frequently encountered problems, a job that only makes sense in a representational setting characterized by asymmetries in our practical and epistemic relations to events. Our representations of Nature are organized around the distinction between what is known and what merely is, and the distinction between what I do and what merely happens. Some beliefs have the role of standing in for features of the world as it appears in our account of Being: the kind of inner proxy for outer objects that is the paradigm of reference in the matching game. But some of them are designed to help us overcome epistemic and practical limitations.[24] Best systems repackage information in forms that are poised to play a role in the kinds of practical and epistemic problems we face. In the case of chance, the right account of their function has to be connected to seeing how they help us resolve opinion in the absence of specific information about an event.[25] And in the case of laws (and other alethic modalities like propensities and capacities and dispositions), the right account has to be connected in a very deep way to seeing how they help resolve decision, where counterfactual implications matter.

To conceive of representation (whether it is the formation of beliefs, or the construction of best systems) as wholly engaged in reflecting features of a landscape that is separate and independently well-defined treats it as a kind of passive reproduction, as though history, from the mind's perspective, were a fixed field of events that simply comes into view in stages. But there is no stable position that sees cognitive activity as in the business of merely reflecting what is "there anyway" or "there already." It views history as partly constituted by its own activity.[26] Our *decisions* about what *to* do make a difference to our *beliefs* about what *we will* do, and our beliefs about what we will do, make a difference to how we expect history to play out. If I haven't decided whether to pin with my rook or castle with my knight, my expectations about what will happen are as open as my decisions. To the extent to which what happens at later stages depends on what happens next, it hangs on my decision. Until a decision is rendered, my expectations about how the game will play out are conditional. There is no way of stabilizing our beliefs about the part of the Humean mosaic that lies in the future without stabilizing our plans. And the ordinary way of stabilizing one's plans involves imagining the results of different potential paths. That is what makes reasoning for an embedded agent – that is, a participant in the events that he represents, not merely an observer – necessarily, in part *practical*. And it is in their role in practical reasoning that makes alethic modalities indispensable for the agent. When he is reasoning about the future, his beliefs about what will happen are inseparable from his beliefs about what he will do. And his beliefs about what he will do depend on his beliefs about what *would* happen under various possible choices for action.

When we adopt the perspective of Being, we abstract from asymmetries in our relations to events around which much ordinary cognition is organized. We abstract not only from the epistemic limitations that give chance its *epistemic* role, we abstract from the practical perspective of a contributor to history that gives the alethic modal beliefs their *practical* role. The distinction between known and unknown and the distinction between what I do and what merely happens introduce asymmetries at the level of content around which many of the structures and concepts that are central to both cognition and science are structured.[27] These give rise to two concomitant notions of modality: the concept of a range of ways the world might actually be so far as I know, and the concept of a range of ways the world could be, only one of which is actual. There are epistemic possibilities and practical possibilities. Belief for an embedded agent is belief about a future that is both uncertain and unsettled. It is uncertain because he has limited information. And it is unsettled because the future is partly constituted by decisions he has yet to make. It is simply not true that every non-logical

term has the job of *standing for* some feature of the world that is "there anyway." If we want an interesting and informative account of epistemic and alethic modality we have to locate modal beliefs in the epistemic and practical context created by the embedded perspective and see what role they play.

Now we can see more clearly both why Lewis failed to provide reductions and why he failed to find useful correspondences. He failed to find reductions because best systems have to have inductive content that can help them resolve uncertainty and decision. And he failed to find useful correspondences because when we adopt a four-dimensionalist perspective and look for static mappings of content-level structures into Being, we abstract from the distinctions that allow us to distinguish the roles that different structures play. There are two sorts of distinctions: distinctions in extension, and differences in internal role. One of the results of the holistic restructuring affected by the BSA is that we won't expect to find differentiation in extension. All of the outputs of a best system reflect the way things generally hang together. We get interesting differentiation among the different structures by looking at the roles they play in our cognitive and epistemic lives.

This doesn't just go for beliefs about laws and chances. Agents form all kinds of belief – beliefs about dogs, colors, democracy, about the value of the American dollar, what tastes good and looks good, what is beautiful, and what is right, what should be eaten, and what should be avoided. Not all of these have the job of standing for something that is there anyway. A fully articulate reproduction of the relationship between belief and Being will exhibit the full variety of relationships between content-level phenomena and Being. There are a lot of distinctions to be drawn in terms of the roles of various structures and the widely distributed facts about world and agent that jointly support those roles. But there is no more compact story about a class of beliefs relates to the world from a God's Eye view than the full dynamic story of the role it plays in the complex two-way coupling between mind and world. The distinctions we draw in the roles different classes of belief play from the side-on are not distinctions between what is *real* and what is not, or what really *refers* and what does not, or between is in *the world* and what is in *us* (as though we weren't in the world). Those dichotomies are all too simple to provide a fully general account of the complex, intertwined coalition of factors concerning how things are with us, on the one hand, and with the world, on the other, that are implicated in even in very simple cases of belief. And they are too simple to capture the array of different roles beliefs play in mediating the interaction between agent and world. These coalitions are revealed in the side-on account of the facts about the world and our situation in it that open up the space for belief in all of its forms.[28]

12.3.1 The Side-On View as a Generalization of the Search for Truthmakers

The side-on account will tell the full story of how such beliefs are formed and used, replacing the expectation of static two-term truthmaker relation between belief-contents and Being with something dynamic and a good deal more complex, something that is both world and agent involving, and opens up a range of differences in direction of fit and cognitive role. The problem with the "every truth has a truthmaker" thesis is that either it interprets the truthmaker relation in a way that tries to assimilates all belief to the paradigm of Price's "matching game," in which case it gets the wrong account of laws and chances. Or it remains neutral about the structure of the truthmaking relation abstracting from differences in cognitive role and direction of fit. In this case, it leaves us with a wholly uninformative mapping of the full body of outputs of theorizing into Being. All theorems of the best system turn out to have the same truthmaker.

Whatever story you tell about how situated agents repackage information for useful deployment in practical and epistemic inferences is going to be complex, dynamic, both agent and world involving. Differentiation between the outputs of the holistic restructuring affected by a theory is going to invoke

differences in *function*, not *extension*. If we want an interesting and informative account of how a class of beliefs finds a place in our view of the world, we have to locate those beliefs in the epistemic and practical context created by the embedded perspective and see what role they play. This means (i) reintroducing the agent as a third term relating model to world, and (ii) taking a dynamic perspective. Representation is thought of not as a relation, but an activity, and an activity pursued by agents. Resituating models in the dynamics that mediates the interaction between agent and environment frees us from an outmoded, a-practical idea of representation is for, and gives us the tools for understanding the complicated ways in which content-level structures relate to Being, and why the idea of truth built on the paradigm of unmediated correspondences between representation and representatum is just not very helpful in this capacity.

12.3.2 The New Humean

My suggestion to the Humean is to distinguish content-level relations from relations between belief and Being. Deny that law and chance beliefs are interchangeable with beliefs about categorical facts at the level of content. Firmly reject extensional analysis of modal belief as "picking out" some feature of the world that is "there anyway"; replace extensional analysis with a side-on account of the role beliefs about laws and chances play in the coupling between agent and world.[29] The side-on account is presented as the generalization of the truthmaker relation. Beliefs are not distinguished just in extension, but internal role. The BSA is presented is presented as part of a side-on view of how beliefs about laws and chances are formed. It is supplemented with a story about the role they play in our cognitive and epistemic lives, facilitating the practical and epistemic inferences.

The Humean agrees with the non-Humean about relations at the level of content. Where the Humean and non-Humean part company is that they have a different side-on view. The anti-Humean thinks that there are modal objects or properties among the fundamental furniture of Being that explain the acquisition and use of modal beliefs in the way that dogs and donuts explain the acquisitions of dog and donut beliefs. The Humean story starts with the BSA and says how beliefs about law and chance package information about the Humean mosaic in forms that are poised to guide belief and action for creatures like us invoking nothing as input to the formation of modal beliefs but local matters of particular, actual fact. What does God see according to the Humean when he views the world *sub specie aeternitatis?* He sees a world that exhibits enough regularity to support scientific activity. He sees the formation of local open subsystems that exhibit behavior robust and regular enough to support modelling. He sees the stabilization of chances and causal pathways at the local level. He sees the emergence of cognitive agents that exploit that regularity to direct behavior to their advantage and how such agents construct models that allow them to gauge the effects of potential interventions in the world. He sees the transformation of the informal construction of models into an inductive art in the hands of science and tells the story of how our expectations under uncertainty and beliefs about the modal substructure of the world become systematized and cast in the form of beliefs about laws and chances much as the BSA said they did. All of this is part of the natural history of Being. And that is all there is.

I will put this by saying that, for the anti-Humean, modal beliefs reflect modal facts. For the Humean, modal facts are shadows of modal beliefs.[30] This changes the site of the dispute so that it is a dispute about the correct account of how modal beliefs are acquired and deployed. This is where I think the dispute should be conducted, and it is where I think the Humean has an advantage. The demand for an intelligible side-on story is the demand to do what Shimony described as "closing the circle," integrating one's ontology with one's epistemology and with one's pragmatics. I haven't tried to fight that battle here for the Humean, but it is the field on which the battle should be fought, and the one on which the Humean has an advantage.

12.3.3 Lewis's Transition from a Deflationary to an Inflationary Modal Ontology

Interestingly, Lewis's own account of belief formation about modal facts – at least the modal beliefs that are the output of scientific theories – doesn't mesh very well with his modal metaphysics. And one of the most important arguments about his modal metaphysics is precisely that it fails the test for providing an intelligible side-on story. Anyone that introduces things into her ontology as truthmakers for modal beliefs owes a side-on account that tells an intelligible story about how our ways of forming modal belief are ways of forming beliefs about *those* things, and why beliefs about *those* things play the role that modal beliefs play in our practical lives. And it is very hard to give such an account if modal beliefs are beliefs about Lewisian possible worlds.

Let us introduce a distinction between inflationary and deflationary realism.[31] The difference between the two is not a difference in the cognitive function or psychological attitude one should take to modal belief. Both are forms of modal realism. The difference between them emerges from the side-on view and has to do with where and how modal beliefs arise. The inflationary realist thinks about A that there is something in the mind-independent fabric of reality, something that is, as I have been saying "there anyway," well defined independently of A-beliefs, and which plays an explanatory role in the production of those beliefs. The deflationary realist tells a story about how A-beliefs arise that doesn't invoke A-facts, and thinks of the A-facts as shadows of A-beliefs. In an inflationary story, the facts about As are part of the input to the formation of A-beliefs. In a deflationary story, the A-facts are part of the output. The deflationist about modal facts denies that the existence of possible worlds, or other kinds of specifically modal truthmakers plays the same sort of explanatory role in the formation of modal beliefs that the existence of trees plays in the formation of tree beliefs, or the existence of dogs plays in the formation of dog beliefs. One gets to be a modal realist without being an *inflationary* realist if one can tell about how modal beliefs get up and running, how they are used, and the role they play in guiding belief and action that doesn't invoke possible worlds, or other kinds of specifically modal truthmakers in an explanatory role.

It is interesting to see the development of Lewis's own views through these lenses. Lewis gave an entirely valid deflationary argument for the existence of possible worlds in *Counterfactuals*:

> I believe, and so do you, that things could have been different in countless ways. But what does this mean? Ordinary language permits the paraphrase: there are many ways things could have been besides the way they actually are. I believe that things could have been different in countless ways; I believe permissible paraphrases of what I believe; taking the paraphrase at its face value, I therefore believe in the existence of entities that might be called "ways things could have been." I prefer to call them "possible worlds."[32]

As far as Lewis says in this passage, possibilities are just ways things might have been. Possible worlds are shadows of beliefs about the way things might have been, leaving open for hermeneutic interpretation beliefs about the way things might have been. But there was a switch between the deflationary argument in *Counterfactuals* and *On the Plurality of Worlds*, where Lewis began defending an inflationary ontology for possible worlds. There Lewis argues that possible worlds are same in kind, different in number from our world. They are distinct, non-overlapping, unified, concrete structured objects.[33]

> When I profess realism about possible worlds, I mean to be taken literally. Possible worlds are what they are, and not some other thing. If asked what sort of thing they are, I cannot give the kind of reply my questioner probably expects: that is, a proposal to reduce possible worlds to something else. I can only ask him to admit that he knows what sort of thing our actual world is, and then explain that possible worlds are more things of *that* sort, differing not in kind but only in what goes on at them.[34]

The reasons he gave were that including possible worlds in our ontology the best account of modal speech involves quantification across possibilities, that our best theory of counterfactuals relies on comparisons between possible worlds, that our best theory of mental and verbal content analyses content in terms of sets of possibilities, and that quantification over properties is best understood as quantification over possibilia.[35] Given the availability of deflationary realism, the question to ask is not whether the right account of these matters quantifies over possible worlds, but whether the world-sized concrete particulars that Lewis introduces as truthmakers for modal beliefs play any explanatory role in those accounts. I don't know the answer to this. One can certainly accept all of the work Lewis did formalizing the semantics for everyday counterfactuals without being committed to his inflationary ontology.

Even among those that reject Lewis-worlds, much of the post-Lewisian discussion of modal metaphysics is organized around the search for truthmakers for modal beliefs.[36] From the point of view of the new breed Humean, this is all scratching where there is no itch. The whole story about laws and chances is given in the side-on account of use.

12.3.4 Objections and Queries

Is this instrumentalism? That depends. I certainly think that laws and chances are tools that help us solve practical and epistemic problems, part of a wider class of structures like dispositions and capacities and causal relations that guide expectation and interaction with a world that is partially known and subject to our input. But I deny some of the traditional accompaniments of instrumentalism, that is, that they are uninterrupted nodes in a formal calculus, that they don't have representational content, or that they are not truth-bearing. In my view, they are structures that are defined on our practical and epistemic interface with Nature, akin to the embellishments we might add to a map of our neighborhood to make it easier to steer by.

Doesn't the fact that the side-on story makes ineliminable use of modal vocabulary show that modal facts have to be part of the fundamental furniture of the world? This is an objection that I've heard several times. Of course we give the side-on account in fully modalized vocabulary. The side-on account is an explanation of how modal belief arises, and all explanations occur at the level of content and presuppose the connexions that are built into that level. I acknowledge that modality is primitive at the level of content. What I deny is that primitive content-level structures correspond to primitive structures at the level of Being. A lot of the structures that define the space within which cognition operates are contingent, emergent, approximate, and perspectival from the point of view of physics. A good many of them are organized around epistemic and practical asymmetries in our relations to events rather than among events themselves. The relationship between content-level structure and structure at the level of Being is mediated by the side-on account which tells us how modal belief arises without inflationary realism.

12.4 Conclusion

I argued that either Humean analyses present themselves as an account of the relations among Ideas, or they present themselves as accounts of the relations between Ideas and Being.[37] If they present themselves as accounts of the relations among Ideas, they fail for one reason. If they present themselves as relations between Ideas and Being, they fail for another. They fail as accounts of the relations among Ideas because there is a separability between beliefs about laws and chances, on the one hand, and beliefs about local matters of particular fact, on the other, in something like the way there is a separability of beliefs about meaning and linguistic behavior or beliefs about the value of a dollar bill

and economic behavior.[38] We can stabilize beliefs about laws and probabilities (regularities distilled out of what we know of the past and projected into the future and epistemically and practically possible situations) because there are regularities that are robust across differences in space and time, and largely indifferent to the results of the individual interactions they are used to guide. They guide our expectations about what does happen, but they are not reducible to claims about what does happen because, in both cases, they have implications for what happens in epistemically and practically possible situations that is indispensable to their cognitive role. If they present themselves as accounts of the relations between Ideas and Being, they fail, because there is no interesting and informative mapping of beliefs about laws and chances into the Humean mosaic.

The lessons to be drawn from this: give up the project of providing content-preserving reductions or mapping Ideas into Being; reconceive the BSA as part of a side-on story about how beliefs about laws and chances are generated. Supplemented with an account of the cognitive and epistemic role they play for situated creatures like us – an explanation of how we form beliefs about laws and chances, and the role those beliefs play guiding expectation and choice – they tell the whole story of how laws and chances earn their place in the world-view of Beings like us.

My advice to the Humean is to dissociate himself from any attempt to provide content-preserving reductions or a semantics for modal belief. Present the BSA as part of a side-on account of the roles that beliefs about laws and chances play in the cognitive and epistemic life of situated agents living in a world constituted wholly by the Humean mosaic. To scratch one's head at the end of the hermeneutic story and insist that there must be something outside the Humean mosaic, that modal beliefs refer to, something that makes them true, something that they represent, is to get bewitched by a picture of *representation* built on the model of Price's "matching game." Wittgenstein once accused philosophers of behaving like little children who, as he said,

> scribble some marks on a piece of paper at random and then ask the grown-up "What's that?" – It happened like this: the grown-up had drawn pictures for the child several times and said "this is a man," "this is a house," etc. And then the child makes some marks too and asks: "what's this then?"[39]

Nowhere is this remark more apt than in the interpretation of the models presented by a scientific theory. It is as though we take a map, embellished with structures we ourselves have developed to guide epistemic and practical reasoning, and then look at it and ask "What is that, then?"

Notes

1 Lewis 1986b, ix.
2 Maudlin 1994.
3 I will focus on laws and chances because they appear in fundamental physical theories and have the most straightforward treatment. What I say will apply also to causes, dispositions, capacities, and potencies.
4 Lewis 1994, 478.
5 That is to say, they ought to pass the logical test for identity: $A \neq B$ iff there are possible worlds in which $(A\&\sim B)$ or $(B\&\sim A)$.
6 Maudlin 2007, 68. The second disjunct can be accepted only on pain of divorcing the notion of physical possibility from the one that science employs.
7 Loewer 1996, 199.
8 Loewer 1996, 198.
9 Albert (n.d.).
10 Of course, in practice, beliefs about the BSA are themselves formed under epistemic constraints; and so we are always simultaneously using information about local matters of particular fact to form opinion about

laws and chances, and using opinion about laws and chances to form particular expectations on the basis of past fact. But this won't affect the discussion here. See Ismael 2008.

11 Loewer recognizes this and fills it out in his version of the BSA. See his remarks about length, etc., in Loewer 1996.

12 We can divide modal notions into two classes: the epistemic modalities which involve the notion of a way the actual world could be, so far as I know, and the alethic modalities which involve the idea of an alternative way the world could have been, *otherwise* than it actually is. Epistemic modalities (e.g. chance and probability) guide belief. Alethic modalities (e.g. cause and law) guide choice.

13 Lewis 1994, 481–2.

14 Lewis 1994, 489

15 A successful reduction would eliminate the need for *specifically* modal beliefs by showing them to be summaries of facts that could be stated in purely categorical vocabulary.

16 "If HS is true, then contingent truths about chance are in the same boat as all other contingent truths: they must be made true, somehow, by the spatiotemporal arrangement of local qualities" (Lewis 1994, 476).

17 Lewis 1994, 476, my emphasis. It is also worth noting here that the "if known" is crucial. I have argued that it makes the revisions to the principle unnecessary. See Ismael 2008.

18 Loewer goes a little farther, expanding the role of chances and laws to include the scientific and explanatory practices in which they figure. In this expanded account, our beliefs and practices form a kind of implicit theory, and, again, the question is what in the world could we assign as reference to law and chance beliefs that would (more or less) satisfy the implicit theory. As he says "[Humean analyses of concepts like law and chance] should be evaluated in terms of how well they ground and illuminate the practices involving the concepts. These practices are reflected in and are to an extent codified by our beliefs involving them" (1996, 183).

19 He writes: "What pattern in the arrangement of qualities makes the chances? In part, features of history up to the moment in question. For the rest, it is the pattern that makes the probabilistic laws, whatever that is" (Lewis 1994, 476).

20 The confusion is, I suspect, a good part of what is concealing the unclarity about Humean reductions.

21 It is misleading, as often remarked, because it uses the notion of "point of view," where what is really wanted is a *non-perspectival* characterization.

22 This is just the generalization of the idea that, as he put it, "every true belief [or every truth statement] is *made true* by some feature of the Humean Mosaic."

23 Price 2007.

24 The notion of the direction of fit commonly invoked to draw some of these differences is not a static one. Nor is it, by itself, enough to draw the distinctions we want to draw among all of the different ways in which a representational state can relate to the world. There is no simple dichotomy that can draw them and the attempts to characterize the notion of direction of fit is testament to that complexity. The notion is usually introduced with a collection of examples – we have belief and memory on one side, and intention and imagination on the other – and attempts to capture the differences (e.g. in terms of direction of counterfactual dependence, normative guiding, or the direction of the controlling conditional intentions, or any such thing) all fail as fully general characterizations of the class of cases, and even if the notion of direction of fit could be characterized to distinguish the menu of examples, it would not be enough to draw all of the differences that need to be drawn among classes of belief. It would fail notably, for example, to provide an account of counterfactual belief.

25 When we adopt the perspective of Being, we abstract not only from the epistemic limitations that give chance its *epistemic* role, we abstract from the practical perspective of a contributor to history that gives the alethic modal beliefs their *practical* role.

26 Even if we distinguish mental activity from physical activity, and restrict history to include only the publicly observable domain of action, from the agent's perspective history bears the marks of her mental activity. Her beliefs about history are not *closed* with respect to beliefs about her own mental activity. She decisions about how to behave, she has beliefs about the effects her decisions will have, and her decisions feed right into her expectations for the future.

27 These asymmetries are reconstructed in relational terms restoring symmetry at the level of Being.

28 Examples of the kind of side-on account I have in mind: Blackburn on ethical beliefs, Callendar on the temporal value asymmetry, Kutach on causation.
29 See Price 2011.
30 This way of putting it captures the fact that the difference is one that is visible only from a dynamic perspective.
31 See Price 2011 and Thomasson forthcoming.
32 Lewis 1973, 84.
33 More explicitly:

1. Possible worlds are just as real as our world;
2. Possible worlds are the same sort of things as our world – they differ in content, not in kind;
3. Possible worlds cannot be reduced to something more basic – they are irreducible entities in their own right.
4. Actuality is not a property but an indexical. Each world is actual relative to itself and merely possible relative to other worlds.

These four claims form the foundation of Lewis's inflationary realism.

34 Lewis 1973, 85.
35 Lewis 1986a, chapter 1, sections 1.2–1.5.
36 See, for example, Adams 1974; Stalnaker 1976; Forrest 1982; Forrest and Armstrong 1984; Sider 2002.
37 I use "Ideas" neutrally to mean the kinds of structured representations that we form mentally or build in science (concepts, maps, models).
38 Beliefs about linguistic meaning are stabilized out of a very complicated set of linguistic interactions and play a role at the individual level guiding behavior. Beliefs about the value of the American dollar are stabilized out of a very complicated set of facts about economic behavior and can play a role at the individual level guiding economic behavior. In each case the separability depends on de facto stability of the former relative to the individual interactions that they guide.
39 Wittgenstein 1978, 17.

References

Adams, R. (1974). Theories of actuality. *Noûs*, 8, 211–31.
Albert, D. (n.d). Physics and chance. Unpublished manuscript.
Forrest, P. (1982). Occam's razor and possible worlds. *Monist*, 65, 456–64.
Forrest, P. & Armstrong, D.M. (1984). An argument against David Lewis's theory of possible worlds. *Australasian Journal of Philosophy*, 62, 164–8.
Hume, D. (1902). *Enquiry concerning human understanding*. Ed. L.A. Selby-Bigge, 2nd ed. Oxford: Clarendon Press.
Ismael, J. (2008). Raid! Dissolving the big bad bug. *Noûs*, 42, 292–307.
Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.
Lewis, D. (1986a). *On the plurality of worlds*. Oxford: Blackwell.
Lewis, D. (1986b). *Philosophical papers*, Vol. II. Oxford: Oxford University Press.
Lewis, D. (1994). Humean supervenience debugged. *Mind*, 103, 473–90.
Loewer, B. (1996). Humean supervenience. *Philosophical Topics*, 24, 101–27. Reprinted in J. Carroll (Ed.) (2004), *Readings on laws of nature*, pp. 176–206. Pittsburgh: University of Pittsburgh Press. Page references are to the latter version.
Maudlin, T. (1994). *Quantum non-locality and relativity*. Oxford: Blackwell.
Maudlin, T. (2007). *The metaphysics within physics*. Oxford: Oxford University Press.
Plantinga, A. (1994). *The nature of necessity*. Oxford: Oxford University Press.

Price, H. (2007). Expressivism, pluralism and representationalism – a new bifurcation thesis, http://sydney.edu.au/time/conferences/epr/price.pdf, accessed May 29, 2014.
Price, H. (2011). *Naturalism without mirrors*. Oxford: Oxford University Press.
Sider, T. (2002). The ersatz pluriverse. *Journal of Philosophy*, 99, 279–315.
Stalnaker, R. (1976). Possible worlds. *Noûs*, 10, 65–75.
Thomasson, A. (forthcoming). *The descent of metaphysics*. Oxford: Oxford University Press.
Wittgenstein, L. (1978). *Culture and value*, 2nd ed. Trans. P. von Wright. Oxford: Blackwell.

13

Where (in Logical Space) Is God?

STEPHANIE R. LEWIS

David Lewis was one of your great correspondents. He kept very nearly every letter he received, and he kept copies of very nearly every letter he wrote. He never did use email, so there are no ephemera in his correspondence to worry about. Everything is there in the file cabinets, and a great deal of it has been scanned. Some of the letters he wrote, and received, were long and detailed discussions and arguments, and some of the ones he wrote amounted to drafts of papers.

The core of David Lewis's philosophy, at least his later work, is his metaphysics. A very great deal of his correspondence, over many years, concerned issues at the core of his (or any analytic philosopher's) metaphysics, that is, analytic philosophy dealing with, to give some examples, the fundamental issues of causation, future contingencies, the truth of counterfactuals, and necessity and possibility. And quite a lot of his correspondence bears, centrally or peripherally, explicitly or not, on the great issues in the philosophy of religion which center around these metaphysical matters.

This chapter falls into six sections:

13.1 Background and Scene-Setting
13.2 The God of the Philosophers
13.3 Free Will
13.4 Divine Evil
13.5 Atonement as Penal Substitution
13.6 The Many-Worlds Theodicy

I do not here discuss a seventh issue on which David had some correspondence: middle knowledge and counterfactuals of freedom.

13.1 Background and Scene-Setting

David wrote two papers squarely on issues in philosophical theology, "Evil for Freedom's Sake,"[1] and "Do We Believe in Penal Substitution?,"[2] left an outline of a third, "Divine Evil," turned from an

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

outline to a paper by Philip Kitcher after David had died,[3] and wrote a fourth with an engaging title, "Anselm and Actuality,"[4] but which does not have in it much attention to philosophical theology. In addition, there are some remarks in a paper mostly about broader issues in metaphysics, "Attitudes De Dicto and De Se,"[5] and some more discussion in the book *On the Plurality of Worlds*.[6]

He was concerned with the problem of evil, and the free will defense to the problem of human-caused evil, over many years. There is a great deal of correspondence on the matter. He also, in his letters, discussed middle knowledge at length with many philosophers. Other letters concern atonement, the coherence of theism, and the justification of horrendous evils as a necessary part of great goodness. Other issues are discussed in the correspondence as well, including some dealt with in his published work either not at all or only briefly.

David was interested in philosophy of religion for its own sake, not just as a specialty within broader analytic metaphysics. He taught undergraduate seminars on philosophical theology and on the problem of evil. Indeed much of his metaphysics, especially his work on counterfactuals, is of consequence for philosophers working in analytic philosophy of religion.

This chapter is part of a project to collect David's correspondence explicitly on topics in the philosophy of religion and arrange major threads by topic, tying the correspondence to his published work. He corresponded with 34 people on matters in philosophical theology.

This chapter confines itself to only a few letters on only a few topics. I think that a few letters set out in full (or nearly, with only irrelevant bits omitted) make for a more interesting discussion than a bunch of snippets. (A Snippet, in case you didn't know, is a species of minor demon, not really dangerous but capable of being very, very annoying. Screwtape had hundreds of them at his command.)

It has in it an exchange between David and Alvin Plantinga on the free will theodicy, a letter from David to Michael Tooley on the possibility of divine evil, and letters to Phil Quinn and to and from Bruce Langtry on atonement as penal substitution. There is also a long interchange with Keith DeRose on knowing the truth values of counterfactuals of freedom.

At the end, the chapter departs from reporting the correspondence, and concludes with a discussion of the place which, say I, God must have in David's logical space and also of the success of the many-worlds theodicy that David briefly discusses.

David was never a theist, and for most of his life was a "contented atheist," as he described himself in a letter to Michael Tooley of July 20, 2000. He had sustained correspondences, sometimes over many years, with a number of his Christian friends, people whom he liked and respected. There were lots of points of agreement with them, either initially or after discussion, and also points of irresoluble disagreement about fundamental issues. Here's what he says in print in "Evil for Freedom's Sake":

> I am an atheist. So you might suspect that my purpose is to debunk free-will theodicy, and every other theodicy besides, so as to provide – at last! – a triumphant knock-down refutation of Christianity. Not so. I am convinced that philosophical debate almost always ends in deadlock, and that this case will be no exception.[1] When I argue that free-will theodicy meets with difficulties, I mean just what I say, no more and no less. I am not saying, and I am not slyly hinting, that these so-called difficulties are really refutations. In fact, I wish free-will theodicy success, or at least some modicum of success. I don't want to have a proof that all the Christians I know are either muddle-heads or devil-worshippers. That conclusion would be as incredible as it is unfriendly. But I won't mind concluding that a Christian must believe one or another of various things that I myself find unbelievable. For of course I knew that all along.

He adds, in a footnote, a further remark about philosophical debate that ends in deadlock:

> 1. That may suggest an "anything goes" attitude toward philosophical questions that I neither hold nor approve of. I would insist that when debate over a philosophical question – say, the question whether I

have hands – ends in deadlock, it does *not* follow that there is no truth of the matter; or that we don't know the truth of the matter; or that we ought to suspend judgement; or that we have no reason for thinking one thing rather than the other.

This is how David did philosophy, and how he conversed and corresponded with those with whom he agreed, more or less, and also with those with whom he could engage over fundamental disagreements.

David had a lot of time for philosophical disagreement, even nose-to-nose dispute, provided that it was all in good faith and all in the aid of the discussion. What he didn't have much time for was people who were presumptuous, or manipulative, or incorrigibly silly. And while he did have time for the more benign sort of ratbag, – he knew perfectly well that he was sometimes thought to be a bit of a ratbag himself – even that patience had its limits. Here is a letter that David wrote in 1998, taking a dim view of bibliomancy:

4 May 1998

Brendan McKay
Department of Computer Science
Australian National University
Canberra
ACT 0200

Dear Brendan McKay,

I much enjoyed the session on bibliomancy that Saul Kripke organized at Princeton. I have two suggestions for further experiments.

Suppose that instead of the Bible, you take a long initial segment of pi and convert pairs of digits into letters: 00 = A, 01 = B, . . . , 25 = Z, 26 = A, . . . 51 = Z, 52 = A . . . 99 = V; or any of many other codings. If message are found, this can be held to confirm Descartes' notion that God somehow created the eternal truths of mathematics.

Suppose that instead of the famous rabbis, you search for messages like ESSENDON, SHEEDY, ON YA DONS, CARN BOMBERS, PIES SUCK, and so on, and so forth. This would address a burning theological question: who does God barrack for? No doubt those of other faiths would get busy finding conflicting messages, helping prove just how abundant these messages really are.

Best wishes,
David Lewis

c: Maya Bar-Hillel
Department of Psychology
Columbia University
New York, NY 10027
Saul Kripke

But the rest of this paper, and indeed the whole of the "David and the Christians" project, is concerned not with this kind of stuff, but rather with philosophical engagement between philosophers who were far more similar, philosophically, than they were different, and who had a common view about how to do philosophy and about what was important.

David says to Bob Adams in 1994:

Dear Bob, 21 October 1994

Thank you for your good comments. It's really nice when philosophers who disagree as much as we do about the main issue can agree so well – perfectly, I think – about the shape of the disagreement. It does mean we won't have much by way of argy-bargy to offer at Notre Dame.

And, in the same spirit, from Peter van Inwagen, in a letter to David, dated December 22, 1981 (and no, Peter did not misrepresent David, either in this letter, or as far as I know, anywhere else.)

> It really does amaze me how much agreement you and I can come to when we're applying possible-worlds language, considering the fact that we don't mean the same thing by it at all.
>
> None of this letter is designed to convince you of anything, of course. You and I are both beyond being convinced by the other – about possible worlds, I mean. But I'd be interested to know if I've misrepresented you on any point.

13.2 The God of the Philosophers

The god that David thought most worth not believing in is a version of the Christian God, or the Judeo-Christian God, called by David, following Antony Kenny,[7] "the God of the philosophers." This God is a being who:

- is eternal;
- is omnipotent;
- is omniscient;
- is perfectly good;
- and in particular loves us all;
- is the creator and sustainer of all besides himself;
- is a rational agent;
- is a being whose nature is accessible to us;
- is a being whose actions are, or anyhow always could be, comprehensible to us;
- can, and perhaps also does, have causal impact on what he has created, and in particular on us.

This set of properties characterizes not just one god, but a whole very large set of them. For example, is God eternal by virtue of being *everlasting*, enduring through all time, or by virtue of being *timeless*, outside of time? If God is everlasting, is it that he sees time just as we do, and time passes for him just as for us? If this is how God is, there are two (at least!) ways to treat divine foreknowledge. On one view, divine foreknowledge really would be *fore*knowledge, and God has to wait and see what happens, same as the rest of us. This makes the problem of reconciling divine foreknowledge of the actions of free agents with the freedom of those agents much easier than for a God whose existence is timeless. On another view of an everlasting God, –that of Alvin Plantinga, for instance– God *nonetheless* has essential foreknowledge of free actions.

Metaphysics has immediate, major consequences for philosophical theology. We knew that. For example, I would very much like to find letters in David's file of correspondence explicitly discussing what sort of God a presentist could believe in.

13.3 Free Will

David's metaphysics sometimes takes the form of philosophical theology, especially in his correspondence about the free will theodicy. Here is an exchange with Alvin Plantinga from 1985.

The first letter is from David to Al, dated October 12, 1985: here is the relevant part of it, followed by an excerpt of a reply dated October 26, 1985 (reprinted by kind permission of Alvin Plantinga).

When we met in Wellington and Melbourne, I wanted to talk to you about Free Will Theodicy, but there seemed to be no very good occasion to talk at length.

I hope you won't mind if, for the sake of brevity, I speak as though you were committed to the truth of FWT. Of course I understand that you are not; the most you are committed to is (1) that FWT might be true, and (2) that if true it would explain how God and evil can coexist. So when I ask "Is it so that X?" take that as short for "Do you think a tenable version of FWT could hold that X?"

FWT holds that God values His creatures' exercise of significant freedom; and rightly so, since the exercise of significant freedom is a very good thing. I ask: is it good only sometimes, namely when the creature chooses rightly? Or is it good always, even when the creature chooses wrongly? If it's good always, then when Stalin or Satan freely chooses to inflict horrendous suffering, we do indeed have some evils, both of sin and of suffering, yet also we have something – the exercise of significant freedom – which is very good. Not *as* good, presumably, as an exercise of significant freedom which turns out right, but still a good of the very sort for the sake of which God permits sin and suffering to go on. If so, it would have detracted from this good if anyone, God or man, had rescued the victims. For then the stakes would have been less, so the freedom exercised by Stalin or Satan would have been less significant. The evil has to be there to confer significance on the choice. If an exercise of significant freedom is a good thing even when evil is chosen, then we have a Moorean "organic unity": the affair is good as a whole, but it has evil parts and it wouldn't have been so good if they had not been there. This version of FWT resembles the aesthetic sort of theodicy, in which the dissonant notes make the whole symphony of world history more glorious than it could have been without them. Is that the right direction for FWT to take?

If not, then I have a different question. God knew which exercises of significant freedom would go which way. We're not to suppose that God could have controlled the counterfactual truths about how any given exercise, if permitted to take place, would turn out – that would be for Him to remove the freedom. But He seems to be in a position to control which ones do and which don't take place. If the ones that would go wrong would not be of value, and would include sin and result in suffering, why doesn't He prevent just those ones? He could control the creatures and take away their freedom in just those cases where He foresees that their exercise of significant freedom would go wrong; or else He could at least rescue the victims, thereby removing not the freedom itself but the significance of the freedom. The only exercises of significant freedom that he'd prevent are the ones that *ex hypothesi* are of no value. The others, the ones that He foresees will go right, He may allow to go forward, entirely uncontrolled. They will be just as much exercises of significant freedom, and will have just as much value, as they would have had if He hadn't prevented the ones that were going to go wrong. Why not?

I think I know how you will answer. If I'm right, our disagreement about counterfactuals takes on theological significance in one more way than I'd previously expected.

Imagine that God has followed the course I suggested. Foreseeing which exercises of significant freedom would go right, He has allowed just those ones to go forward uncontrolled and prevented all the rest. Fred's choice was one of the ones that God foresaw would go right. So long ago (or outside of time) God made up His mind firmly to exercise no control over Fred; and also not to rescue the victims in case Fred chose wrongly. Whatever may happen, God will stand firmly by His decision. Now the time has come for Fred to choose. *Has* God left him free? I say yes. It's settled long ago (or outside of time) that God will neither control him nor detract from the significance of his choice by rescuing victims. When Fred chooses right, God will be unsurprised. If he chose wrong, God would be astounded – but would still stand by His decision not to suddenly put Fred under control or rescue victims. Fred and his surroundings, and God's intentions toward him, are exactly as they would have been in the comparison case in which God allows all the exercises of significant freedom to go forward, no matter how he foresaw that they would go. So Fred is no less free than in that comparison case.

I said that if Fred chose wrong, God would be astounded, but would stand by His decision not to interfere. But instead one might back-track, and I think from other discussions that you would. I think you might say that if Fred chose wrong, God would have foreseen that; so *ex hypothesi* God would *not* have permitted Fred's exercise of significant freedom to go forward, but rather would have made up His mind long ago (or outside of time) to prevent it; so if Fred had freely chosen wrong, then Fred would not have freely chosen wrong; which means that Fred could not have freely chosen wrong; which means that, after

all, God did not leave Fred free. By following the policy He did when He allocated significant freedom to some, and control or insignificance to others, He did after all take away Fred's significant freedom; although He did not interfere in any direct way. I think you might say that by taking my misguided advice, God not only took away the worthless exercises of significant freedom that were going to go wrong, but also took away the valuable ones that were going to go right.

As I think you'll expect, I say that there are two senses, or two resolutions of vagueness, of the relevant counterfactual. In the sense that matters for freedom, I say that if Fred chose wrong, God would be astounded but would not interfere, and so Fred would have exercised his significant freedom in choosing wrong. I'll give you a *different* sense, in which it's true that if Fred chose wrong, God would have foreseen it and followed His policy of not permitting the exercise of significant freedom in such a case; but I deny that this different sense is relevant to whether Fred has been left free. This brings us back to the disagreement we have discussed in connection with Newcomb's problem, and with the vulnerable ants.

Yours,
David Lewis

cc (pages 3–4 only): John Bishop

Now as to the second part of your letter: I find the issues here extremely interesting and am eager to work on them. Unfortunately I have had no time to do so; I have been trying to finish a paper (for a Central division symposium) that is already a month overdue. Furthermore I won't have any time during the next two weeks; I'm trying to finish another paper for the interAmerican conference in Guadalajara. So I'll write you about these matters in about three weeks.

In the meantime, however, I have a couple of questions.

First, let me see if I understand you. You speak of resolving an ambiguity in a certain counterfactual: "there are two senses, or two resolutions of vagueness, of the relevant counterfactual." Which counterfactual is this? Presumably

(1) If Fred were to go wrong, God would interfere

or its contrary

(2) if Fred were to go wrong, God would not interfere.

Now why do you say that there are really two senses of (1)? Do you mean to say that prior to our resolving the vagueness of (1) it has one sense while after the resolution of the vagueness it has two senses, neither identical with the original? What would the two senses be? Are there at least two propositions the sentence (1) could or would express after the resolution of vagueness? Is there some way of specifying or stating these counterfactuals (in other terms)? I'm not clear how you are thinking of this. Do you see resolving the vagueness involved as a matter of or something like a matter of adding conditions to the antecedent? So that under one resolution of the vagueness involved, (2) expresses what can be put more explicitly by something like

(2*) If Fred were to choose wrong, and God were to act on his intention of not interfering with Fred, God would be astounded but not interfere

which is true, while under another what it expresses is something like

(1)** If Fred were to go wrong and God were to follow his policy of interfering with those he has seen will go wrong, then God would not interfere

which is false? How are you thinking of this matter? And why would you think (if you do) that the truth of (1*) is relevant to the question of whether Fred is free, while the falsehood of (1**) is not?

> Further: suppose we state the hypothesis in question: that God has and always has had a policy of interfering with those whom he foresees will go wrong and not interfering with those whom he foresees will go right (and would have had this policy no matter what Fred does), that he has foreseen that Fred will go right, and that he will not interfere with Fred; and suppose we add standard views about God (e.g., that he exists and is omniscient in at least all the worlds in which Fred exists, that he cannot change his mind, and the like). Won't there then be a straightforward argument from the hypothesis to the truth of (1) and hence to the falsehood of (2)? And wouldn't that be sufficient for showing that, under the hypothesis in question, it is not within Fred's power to do what is wrong, so that he does not have significant freedom?
>
> I'm sorry that I can't work on these matters for the next couple of weeks, but as soon as I can, I will. In the meantime I'd certainly like to hear what you have to say in response to the above queries.
>
> Yours,

There are many other discussions in David's correspondence of matters concerning free will. One such is an exchange with Edwin Curley, about the truth of counterfactuals of freedom. Do these counterfactuals, of form "If it were the case that p, it would be the case that q" have truth values? Curley says, as quoted by David,

> . . . if indeterminism is correct, then for any possible world in which the agent freely chooses to do A, there is another possible world just like it, as regards its prior history and laws, in which the agent freely chooses not to do A. Indeterminism guarantees that the truth conditions for counterfactuals of freedom cannot be satisfied.[8]

In this discussion they agree that the worlds that we are concerned with are worlds which are the same up to the occasion of the indeterministic action, and differ thereafter. Future similarities don't count.

13.4 Divine Evil

Now we turn to the next of our selections of David's correspondence on topics in the philosophy of religion: a letter on evil, specifically on the possibility of divine evil, evil perpetrated by God. This is one of David's letters that amounted to the zero-th draft of a paper.

At what turned out to be very near the end of his life, David outlined a third paper in the philosophy of religion, "Divine Evil."[9] This letter to Michael Tooley contains the substance of that outline and the argument it contains. Whether or not it succeeds as a step towards a knockdown, drag-out argument, an atheodicy putting all the theodicists that ever there were out of business once and for all, it is "as firm an anti-Christian manifesto as you could have."[10]

Princeton University

Department of Philosophy
1879 Hall
Princeton, New Jersey 08544-1006 U.S.A.
(609) 258-4289
FAX: (609) 258-1502

20 July 2000

Prof. Michael Tooley
Department of Philosophy
University of Colorado
Boulder
CO 80309-0232

Dear Michael,

I've come to think that the much-discussed problem of evil that we've both written about is something of a sideshow, and that the *real* problem of evil (for Christians, anyway) lies elsewhere.

Mind you, I still think the much-discussed problem of evil is a good objection to theism (unless one is such a sucker as to accept Plantinga's challenge to prosecute the case without benefit of any auxiliary premises except logical truths!) The only good way to dodge it is to doubt that the God of scripture, as opposed to the God of the philosophers, is really omnipotent, but that is more than a little bit heretical.

The much-discussed problem concerns the evil, both moral and natural, that God forbears to prevent. The real problem, I think, is the evil that Christians ascribe to God Himself. To punish insubordination with eternal torment would be, I take it, monstrously unjust, because the punishment would be entirely disproportionate to the crime. To do such a thing, or even to be disposed to do such a thing, is horrendously evil. It also flatly contradicts the contention that God loves His creatures.

It is not the omnipotence of the God of the philosophers that sets up the problem, but rather a central feature of Christianity: the redemption. Without the redemption, it wouldn't be Christianity. (Vague theism and uplift joined to a reverence for the life and teachings of Jesus, such as we might hear from Quakers or Unitarians, is not yet Christianity!) And without something pretty horrendous for us to be redeemed *from*, it wouldn't be the redemption.

The Christian universalists tell us that in the end, nobody will be damned. But is this because God will in the end be merciful and abandon His disposition to punish insubordination? Or is it because in the end, everyone will knuckle under? I think it's meant to be the latter. And if it is, God retains His horrendously evil disposition, even if He never has occasion to manifest it.

Sophisticated Christians may say that damnation doesn't really consist of eternal torment. Rather, it's a state for which eternal torment is an apt (or an inadequate) metaphor. I reply that if God is disposed to punish insubordination by inflicting a state for which eternal torment is an apt (or an inadequate) metaphor, that is no less unjust and evil than if the punishment had been eternal torment itself.

Sophisticated Christians may say that damnation is not a punishment inflicted by God, rather it is a state of alienation from God which the damned choose for themselves. (And perhaps they have the option at any time of ceasing to choose it and thereby ceasing to be damned.) I reply in two parts. First, as a contented atheist I can testify from personal experience that the state of alienation from God is *not* a state for which torment is an apt (or inadequate) metaphor. Second, if this state is so horrible, why doesn't God prevent His creatures from falling into it? To this the answer will be that He respects their free will. This now returns us to one aspect of the much-discussed problem. Anyway, I again reply in two parts. First, I reject the value judgement that incompatabilist freedom is worth so much. (And compatibilist freedom wouldn't prevent God from keeping His creatures out of trouble.) Second, it seems to me that God could have done plenty more than He does by way of gentle urging and luring without thereby removing the incompatabilist freedom of His creatures.

I end with a practical problem. If the God whom Christians worship is horrendously evil, what are we to say about the Christians themselves? Do they not share in divine evil by endorsing it? Are they devil–worshipers? I would like to go on thinking that I wholeheartedly respect and admire certain of my Christian friends, but I'm at a loss to see how that's possible.

Yours,
David Lewis

c: mj

13.5 Atonement as Penal Substitution

Now we turn to another topic: penal substitution. If one person volunteers (or is compelled) to be punished for another's offense, is that really punishment, and does it ever "render it permissible (or even obligatory) to leave the offender unpunished?"[11]

David treats the Atonement as an instance of penal substitution. It has to be said that David, normally pretty ahistorical, is especially so in this case. It also has to be said that he treats the crucifixion as an instance of punishment. So far as I can tell, neither his published works nor his correspondence further considers the question of whether atonement is just another kind of punishment, or whether there is something that sets atonement apart from the kind of punishment visited on criminals or wayward children.

A letter from David to Phil Quinn states the issue:

Prof. Philip Quinn 26 February 1996
Department of Philosophy
University of Notre Dame
Notre Dame, IN 46556

Dear Philip,

I recently came upon your paper "The Traditional Understanding of the Atonement Must Be Modified" reprinted in an OUP book of readings. (The book where it appeared originally is not in our library. Can you tell me, please, whether OUP made any cuts or editorial alterations? Or whether they changed the title, as they seem to have done in other cases? Thanks.)

I've had in mind for a year or so that I would write a paper titled "Do We Believe in Penal Substitution?" The first half of my paper would have been redundant given yours, so it's just as well I never got around to writing it! I would have said, of course, that the leading Christian theory of the Atonement conflicts with what most of us – Christians and others alike – believe about just punishment in other sorts of cases.

My second half, however, might still be worth writing – though now it might best be written as a comment upon your paper. My point would be that it's not just Christians (some of them) who seem to be of two minds about penal substitution. More or less nobody thinks that a beneficent substitute can serve my prison sentence, or my death sentence; but more or less everyone does think that a beneficent substitute can pay my fine. If I'm sentenced to pay a fine, and you pay it for me, then it's been paid and I no longer owe it. Why this difference between fines and other punishments?

(1) Because fines are such small punishments? Well, often they are, but sometimes they aren't. Some fines are a more severe punishment than some prison sentences. Yet it seems to me that my opinion that penal substitution is OK in the case of fines applies to large and small fines alike. Or (2) because fines are really just fees, and carry no disgrace? Well, sometimes they don't, but sometimes they do. Yet it seems to me that my opinion applies alike to fines that are really just fees and to fines that carry disgrace. Or (3) because it's impractical to prevent you from paying my fine if you really want to? (We could scarcely run a system on which, if I'm sentenced to pay a fine, you are forbidden to give me a gift equal in value to the fine I must pay!) True; but if we whole-heartedly disbelieved in penal substitution, we ought to take that as an argument that fines are not a satisfactory punishment. Yet I think I've never heard any such argument made.

And, in addition there is an interesting interchange with Bruce Langtry on atonement as penal substitution (reprinted by kind permission of Bruce Langtry).

Dear David, August 6, 1996

Thanks for your draft paper. I have only one comment: perhaps it would be more complete if it contained some discussion of substitution in collective punishment.

Sometimes a teacher detains a whole class after school, as a punishment for its misbehaviour. There are three categories of individual: (i) Those who participated directly and fully in the offence for which the punishment was imposed – eg., shouting and overturning desks (ii) Those who participated indirectly – they did not shout or overturn desks, but they did contribute to the unruly atmosphere immediately leading up to the main offences, or at least they contributed to the background unruly attitude of the class over preceding days, whether or not they were there at the time of the main offences (iii) Those who were completely uninvolved – eg., they are always well-behaved in class. I am inclined to think that the punish-

ment of the individuals in category (iii) is unjust, but punishment [o]f individuals in category (ii) sometimes just. At any rate, the chief point is that the punishment is collective.

The Roman army is said to have sometimes punished seriously misbehaving regiments or other military units collectively, by selecting 1 man in every 10 by lot, and executing those selected. The idea was not that the individual survivors were lucky enough to escape punishment; rather they participated in the punishment imposed on the whole regiment via the execution of those selected.

Suppose that we agree that this method of punishing a regiment was just – at least if it involved the execution of only individuals in categories (i) and (ii). Suppose now that one of the individuals who is not chosen by lot offers to take the place of his friend, who was chosen by lot to be executed. Would this strike the authorities as senseless? Would it undermine the functions of the punishment (according to the retributivist, expressivist, or deterrence theories, etc)? I think not.

Suppose further that the friend who volunteered was in category (iii) – eg., he was a member of the regiment, but had in no way contributed to the offence committed by the regiment. Here is a case in which an innocent person is executed in place of one who shared directly in the collective guilt.

There are many considerations which might lead the authorities to accept the offer. It is not clear that in doing so they would be acting unjustly or pointlessly. To the extent that you agree, you have intuitions favouring penal substitution which go well beyond the area of big fines.

One point made by some Christians is that Christ, while innocent because he had done nothing wrong, nevertheless identified himself with human beings collectively. He became a member of the human community. He was therefore in a position to volunteer to take upon himself the full penalty which the human race collectively deserved, and thereby to save individuals from the punishments to which they were (distributively) liable.

I don't want to develop this line of thought further; indeed I don't quite know how to. But it seems well worth considering in the context of your paper.

Regards,
Bruce Langtry

Here is David's reply:

7 August 1986

Dear Bruce,
Thank you for your letter.

I've written elsewhere about decimation of a mutinous regiment: "The Punishment that Leaves Something to Chance", Philosophy and Public Affairs, 1989, 53ff. I take the line that it is not a collective but an individual punishment: that it is a probabilistic punishment, a "penal lottery". "Each soldier is punished for his part in the mutiny by a one-in-ten chance of being put to death." p. 58. I ague that penal lotteries are unjust; although there's a sense in which the equally guilty are punished equally – their risks are equal – pretty clearly that's not the right sense.

You ask me to suppose, counterfactually, as I think, that decimation is just. Then would allowing substitution detract from its supposed justice? If forced to answer, I say yes. But I'd rather not answer. Because the initial supposition is so far-fetched, I lose control of it. Too much is up for grabs. Some of my moral opinions are to be held fixed, some suspended, and which are which? (Besides, maybe I'm supposed to imagine myself thinking as an ancient Roman would, and I don't know enough about how ancient Romans thought.[)]

[Bruce's Melbourne Uni colleague Allen] Hazen made a different connection between penal substitution and penal lotteries. Insofar as it's a matter of luck whether the convicted offender can or can't find a willing substitute, a system that tolerates penal substitution is to that extent a penal lottery. And insofar as it's not a matter of luck but not a matter of desert either, that's if anything even worse: the rock star has his fans, he will surely find a substitute and go free: not so the logic lecturer.

Yours,
David

13.6 The Many-Worlds Theodicy

David considered a theodicy that only a modal realist can even attempt to run. (Peter Forrest once said that David was the most extreme modal realist going.) David says[12] to his old friend Jonathan Bennett:

> I do think there are sometimes moves in analytic philosophy of religion that may look good to a puzzle-solving philosopher but are not religiously serious. Take, for instance, many-worlds theodicy:
>
> Why didn't God make the best possible world? – He did. Whaddaya mean, best world? Lisbon Earthquake, AIDS, Pol Pot, I said God made the best possible world. I never said this was it. God made the best, He made the second-best, He made the third best, and so on down through all the worlds with a positive balance. We're down there somewhere.
>
> I think this is a nifty solution to a philosophical problem (but not so nifty that it clearly works – maybe it plays on mixing two different notions of what it is to be a world). But I doubt that it could really be part of someone's religion. I'm not sure why not, and I could use some help from native informants. I know of three inventors, – independent so far as I know, – of many-worlds theodicy. Burnheim[13], radicalized ex-priest, thought it was a joke (conversation, Sydney Harbour, 1971). McHarry[14], who I know nothing about, writes as if in dead earnest (Analysis, 1978). With Nozick (Examined Life, 227 ff.) you can never tell.

Now here's what I think. I think that this theodicy can't be made to work in David's modal realism. Aside: it doesn't even begin to work in ersatz modal realism, in which the possible worlds – ersatz possible worlds – are made out of components of this world. In ersatz modal realism there really is only one possible world, this one. And there is (at most) one God, and the counterexamples to the premise that this is the best of all possible worlds bite just as hard as they do in the traditional atheodicies that turn on the presence of evil in this world.

I think David doesn't get to the real reason why his many-worlds theodicy doesn't work. This is a huge surprise to me: David always got to the real reasons. The problem isn't just that "maybe it [the many-worlds theodicy] plays on mixing two different notions of what it is to be a world." It's worse than that.

Why can't David run the many-worlds theodicy? For utterly fundamental reasons built into his modal realism.

For starters, let's look at David's description of our world:[14]

> The world we live in is a very inclusive thing. Every stick and every stone you have ever seen is part of it. And so are you and I. And so are the planet Earth, the solar system, the entire Milky Way, the remote galaxies we see through telescopes, and (if there are such things) all the bits of empty space between the stars and galaxies. There is nothing so far away from us as not to be part of our world. Anything at any distance at all is to be included. Likewise the world is inclusive in time. No long-gone ancient Romans, no long-gone pterodactyls, no long-gone primordial clouds of plasma are too far in the past, nor are the dead dark stars too far in the future, to be part of this same world. Maybe, as I myself think, the world is a big physical object; or maybe some parts of it are entelechies or spirits or auras or deities or other things unknown to physics. But nothing is so alien in kind as not to be part of our world, provided only that it does exist at some distance and direction from here, or at some time before or after or simultaneous with now.[15]

What about other possible worlds? David says:

> . . . absolutely every way that a world could possibly be is a way that some world is.[16]

Where in logical space is God? David says perfectly explicitly that God, if such there be, is a part of a possible world.

> I too might be heard to say that there is only one world, that there are only finitely many donkeys, or that there is no God. But when I say such things, I am restricting my quantifiers, just as when I look in the fridge and say that there is no beer. I do not deny that there is beer outside the fridge, but I ignore it in my speech. Likewise I may ignore the other worlds, and the other-worldly donkeys and gods, without at all denying that – speaking unrestrictedly – they do exist.[17]

He also says:

> The worlds are not of our own making. It may happen that one part of a world makes other parts, as we do; and as other-worldly gods and demiurges do on a grander scale. But if worlds are causally isolated, nothing outside a world ever makes a world; and nothing inside makes the whole of a world, for that would be an impossible kind of self-causation.[18]

Before we get down to work on the many-worlds theodicy, let me first clear away a distraction. There is another, completely distinct, set of arguments, the *no-best-world* arguments, supporting the conclusion that God could not have created the best of all possible worlds, because there isn't one. Robert Elliott[19] and several others take the line that there is no best of all possible worlds: for every world there is a better one. So God, try as He might, could not create a possible world that is the best of all possible worlds.

Some versions of this argument, as do those just mentioned, turn on an argument that the ranking of possible worlds as ordered by goodness has no upper limit: for every possible world there is a better one. Another, and distinct, line of argument, which I do not discuss here, could turn on a premise that goodness consists of too many properties to allow for a single, well-ordered ranking of worlds as ordered by their goodness.

Still other no-best-world arguments, including Bradley Monton's,[20] rely on the truth of the claim that God can create many instances of qualitatively identical worlds. If God can do this, then there is no unique best of all possible worlds, though there might be a set of worlds, identical with one another in all respects, that rank highest in goodness.

The no-best-world atheodicy has legs. Bradley Monton's paper is recent, and he cites several other recent discussions of the no-best-world atheodicy. The discussion continues. But the no-best-world arguments, all of them, differ fundamentally from the many-worlds theodicy that David describes in his letter to Jonathan Bennett. The no-best-world arguments rely on the success of an argument that there just *is* no best of all possible worlds. Arguments of this sort can be made against any theodicy, or indeed any atheodicy, with a premise in it that says that God created, or must create, the best of all possible worlds. This is entirely independent of any theory about modal realism, or ersatz modal realism, or the view that there is exactly one world, namely, this one.

Now, back to business, and the many-worlds theodicy that David talks about in his letter to Bennett. There doesn't seem to be much discussion of this in the literature.

The McHarry paper is the best statement, all three pages of it, that I know of, of the argument. McHarry considers a world – a possible world – to be a collection of all things spatiotemporally related to one another. He and David agree about this. Two such disjoint collections do not constitute one world. He and David agree about this too, and indeed this is fundamental to David's treatment of possible worlds as causally insulated from one another. According to McHarry, on the assumption that there could be more than one such possible world, then the problem of evil dissolves, since God can, indeed must, actualize every world which is better than nothing. This is David's argument-sketch.

This is a stronger argument than David's. David has it that there *are* all those possible worlds. McHarry says only that if there *could be* possible worlds, then the theodicy works.

Now the troubles begin. First let's look at a little trouble, then at the big one.

The little trouble arises from David's claim, stated above, that nothing inside a world ever makes the whole of that world, for that would be an impossible kind of self-causation. This is a problem all right, but a problem not unique to modal realism that gives us a pluriverse of possible worlds.

Indeed, it looks as if David is not entirely committed to this claim. In the interchange with Ed Curley mentioned above, David says:

> I say: there is an actual world. To be sure, most of it lies off in the future of the moment prior to the creation. There are truths about what will in fact happen, even though most of those truths are not yet predetermined. Only if we've made a case that similarities involving actual future contingencies don't count (and I don't object at all to that view, I just say it needs to be argued), or if instead we doubt that there are present truths about future contingencies, can we say that there's nothing yet to settle the truth of the conditional.[21]

We have a divergence between the God of the philosophers and the gods that are located in some, but not all, possible worlds. The God of the philosophers is the creator, and perhaps also the conservator, of all besides himself. If we restrict our quantifiers, as David would have us do, we can say that the God *of any possible world* is the creator of all the rest of *that* world, and only that world. But in order for this to work we have to abandon David's view about impossible kinds of self-causation. What happens if we do abandon this? Nothing very drastic. As a change to David's modal realism, this is pretty cheap. It loses nothing about the plurality of worlds, about the mechanics of the counterpart relation, or about the claim that some possible worlds have gods in them and some don't.

What *would* be drastic is to have to say that there is (at most) one and only God, the God of all possible worlds (or of none), not a member of any one possible world. And here, as Captain Hook[22] said in the musical *Peter Pan*, is where the canker gnaws. I think that to run the many-worlds theodicy as David states it you need a single and unique God, one God, not a part of any world, not itself a world, but a third kind of thing, an entity all by itself. This kind of God looks like the God of the philosophers. But to have that status it must be an entity which has causal impact on every possible world.

Well, what if we try adding to logical space a single being, God, which is not the same sort of thing as a possible world, is a part of no possible world and, uniquely, has causal impact on *all* of them?

But Ludovician possible worlds are causally insulated from one another, with no common members, and are such that nothing has causal impact on more than one world. Each world has its contents, and nothing that is in one world is in another. Something in one world may have *counterparts* elsewhere, but no individual can be in more than one possible world. This is what makes them possible worlds, distinct from one another, and is fundamental to David's mad-dog modal realism.[23] To say that something out there in logical space can have causal impact on *all* the possible worlds, indeed, created them, but is itself a member of no single world, is profoundly, fundamentally, completely, and utterly inconsistent with David's modal realist metaphysics of possible worlds.

Notes

1 *Philosophical Papers*, 22 (1993), 149–72; reprinted in David Lewis, *Papers in Ethics and Social Philosophy* (Cambridge: Cambridge University Press, 2000).

2 *Philosophical Papers*, 26 (1997), 203–9; reprinted in David Lewis, *Papers in Ethics and Social Philosophy* (Cambridge: Cambridge University Press, 2000).

3 In Louise M. Antony (Ed.), *Philosophers without Gods: Meditations on Atheism and the Secular Life* (Oxford: Oxford University Press, 2007).
4 *Noûs*, 4 (1970), 175–88. Reprinted with added postscripts in David Lewis, *Philosophical Papers*, Vol. I (Oxford: Oxford University Press, 1983).
5 *Philosophical Review*, 88 (1979), 513–43. Reprinted with added postscripts in David Lewis, *Philosophical Papers*, Vol. I (Oxford: Oxford University Press, 1983).
6 Oxford: Basil Blackwell, 1986, reissued by Blackwell, 2001.
7 Oxford: Clarendon Press, 1979.
8 In a draft of a paper "The Incoherence of Christian Theism."
9 "Divine Evil." In Louise M. Antony (Ed.), *Philosophers without Gods: Meditations on Atheism and the Secular Life* (Oxford: Oxford University Press, 2007). David wrote only an outline; Philip Kitcher fleshed out the paper for publication.
10 Richard Swinburne's phrase. Swinburne here combines disagreement and respect.
11 "Do We Believe in Penal Substitution?" *Philosophical Papers*, 26 (1997), 203–9. Reprinted in David Lewis, *Papers in Ethics and Social Philosophy* (Cambridge: Cambridge University Press, 2000).
12 In a letter to him dated May 26, 1992.
13 This is John Burnheim, of the philosophy department at Sydney University. Burnheim himself seems not to have published on the matter, and there is no mention of it in his correspondence with David.
14 *On the Plurality of Worlds* (Oxford: Basil Blackwell, 1986), 1.
15 It's a pity that David didn't include Heraplem the lollipop, its stick Paraplete, its redness Harlac, and its roundness Hamis. All this comes from Donald Cary Williams, "The Elements of Being," *Review of Metaphysics*, 7 (2) (1953), 3–18, 171–92. David helped to get Williams's tropes back into play, after years of undeserved neglect.
16 *On the Plurality of Worlds*, 2.
17 *On the Plurality of Worlds*, 136.
18 *On the Plurality of Worlds*, 138.
19 "Divine Perfection, Axiology and the No Best World Defence," *Religious Studies*, 29 (1993), 533–42.
20 "Against Multiverse Theodicies," *Philo*, 13 (2010), 113–35.
21 David to Ed Curley, March 3, 2000.
22 Played by the great Cyril Ritchard.
23 Bill Lycan's way of putting it, in his review of Lewis's *On the Plurality of Worlds: Journal of Philosophy*, 85 (1988), 42–7.

14

De Re Modality, Essentialism, and Lewis's Humeanism

HELEN BEEBEE AND FRASER MACBRIDE

14.1 Introduction

Modality is standardly thought to come in two varieties: *de dicto* and *de re*. While it is a *de dicto* necessary truth that all bachelors are unmarried, it is not (we may suppose) true of a particular bachelor – call him Jack – that *he* is necessarily unmarried; for surely Jack – that very same particular man – *could* have been married. By contrast, we might hold that it *is* true of Jack that he is necessarily human: he could not have been the very object that he is and yet failed to have been human. *De re* modality concerns the attribution of modal features to *things* or *individuals*.

W.V. Quine was notoriously skeptical about *de re* modality, holding that it enshrines a commitment to Aristotelian essentialism – a position he professed to find utterly bewildering. In section 14.2 of this chapter, we explain how Lewis's earliest (1968) account of *de re* modality is an attempt to answer Quinean skepticism about *de re* modality. However, we argue in section 14.3 that his later, developed view, according to which facts about essences are context-relative, provides a decidedly skeptical solution to Quine's skeptical problem. As Lewis himself puts it: "The true-hearted essentialist might well think me a false friend, a Quinean skeptic in essentialist's clothing" (1983a, 42). For example, whether or not Jack is essentially human depends upon the context within which we are asking the question: there is no absolute, context-independent fact of the matter.

In sections 14.4–14.6, we consider how Lewis's conception of *de re* modality fits into his overall metaphysics. Our hypothesis is that the driving force behind his metaphysics in general, and his adherence to counterpart theory in particular, is the distinctly Humean thought that necessary connections between distinct existences are literally unintelligible. In section 14.4, we show how Lewis's appeal to counterpart theory in his account of truthmakers is explicitly aimed at delivering a truthmaker principle that eschews necessary connections. In section 14.5, we argue that Lewis's attitude towards several well-rehearsed debates in contemporary ontology – and the reasons underlying that attitude – closely mirror Hume's incendiary verdict on "divinity and school metaphysics." Finally, in section 14.6, we connect his account of *de re* modality with his adherence to the doctrine of "Humean supervenience," and argue that his adherence to a Humean world-view runs considerably deeper

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

than mere commitment to the contingent truth of this thesis. Rather, Lewis upholds Hume's insight that necessary connections between distinct existences are unintelligible, and not merely alien to worlds like ours. Lewis's counterpart theory is one of the key elements of his overall metaphysics that allows him to uphold this insight.

14.2 *De Re* Modality and Counterpart Theory

Quine's avowed bewilderment about *de re* modality is expressed in the context of his hostility to quantified modal logic (QML): a modal logic that deploys the box (necessity) and diamond (possibility) operators within the scope of quantifiers (see e.g. Quine 1963). Thus while "□ (all mathematicians are rational)" can be seen as a relatively harmless expression of *de dicto* necessity, a sentence such as "(∃x) □ (x is rational)" must, if it is to make any sense at all, be given a *de re* reading. But Quine simply cannot see how to make any sense of such a *de re* reading, and concludes that QML is fatally flawed.

What, then, is wrong with *de re* modality, according to Quine? the answer is that he takes it to enshrine a commitment to what he calls "Aristotelian essentialism":[1] the doctrine that

> [a]n object, of itself and by whatever name or none, . . . [has] some of its traits necessarily and others contingently, despite the fact that the latter traits follow just as analytically from some ways of specifying the object as the former traits do from other ways of specifying it. (1963, 155)

And he has no time at all for Aristotelian essentialism; as he says,

> the way to do modal logic, if at all, is to accept Aristotelian essentialism. To defend Aristotelian essentialism, however, is not part of my plan. Such a philosophy is as unreasonable by my lights as it is by Carnap's or [C.I.] Lewis's. And in conclusion I say, as Carnap and Lewis have not: so much the worse for quantified modal logic. (1963, 156)[2]

And why the hostility to Aristotelian essentialism? Quine sheds some light on this question in *Word and Object*:

> Perhaps I can evoke the appropriate sense of bewilderment as follows. Mathematicians may conceivably be said to be necessarily rational and not necessarily two-legged; and cyclists necessarily two-legged and not necessarily rational. But what of an individual who counts among his eccentricities both mathematics and cycling? Is this concrete individual necessarily rational and contingently two-legged or vice versa? Just insofar as we are talking referentially of the object, with no special bias towards a background grouping of mathematicians as against cyclists, or vice versa, there is no semblance of sense in rating some of his attributes as necessary, and other as contingent. Some of his attributes count as important and others as unimportant, yes, some as enduring and others as fleeting; but none as necessary or contingent. (1960, 199)

Quine's basic complaint, then, is that Aristotelian essentialism requires us to talk "referentially of the object, with no special bias towards a background grouping." To put it another way, Aristotelian essentialism commits us to *de re* modality: it commits us to thinking that a given object *x* is such that *it* is necessarily or contingently thus-and-so. And Quine is simply bewildered by this thought: it contains "no semblance of sense." Consider Gavin, the cycling mathematician. Is *he* necessarily rational and only accidentally bipedal, or rather necessarily bipedal and only accidentally rational? Or is he, perhaps, necessarily both rational and bipedal? If Aristotelian essentialism is to make any sense then these questions must have determinate answers. The only possible way of giving them determinate answers, Quine thinks, is by adopting a "special bias towards a background grouping"; and that, Quine thinks, is incompatible with talking purely referentially of the object, which is what is forced on us by Aristotelian essentialism and hence by QML.

In his "Counterpart Theory and Quantified Modal Logic" (1968), Lewis offers a radical alternative to QML. Whilst QML adds modal operators – the box (necessity) and diamond (possibility) – to our existing first-order logic, Lewis's "counterpart theory" simply extends the domain of quantification of first-order logic, so that it quantifies over not just the actual world and its inhabitants but all possible worlds and their inhabitants – related by the counterpart relation. The result is a logic into which the sentences of box-and-diamond modal logic can be translated, but which is stronger than QML (since there are sentences of counterpart theory that cannot be translated into QML).

Counterpart theory is constructed out of standard first-order logic by the introduction of four primitive predicates (*x is a possible world*, *x is in possible world y*, *x is actual*, and *x is a counterpart of y*) whose behavior is governed by a set of axioms ("nothing is in anything except a world"; "nothing is a counterpart of anything else in its world"; "some world contains all and only actual things"; etc.; Lewis 1968, 113–14). The translation from QML into counterpart theory takes sentences deploying boxes and diamonds and translates them into the language of counterpart theory, so that "$\Diamond Fa$" translates as "*F* holds of some counterpart *c* of *a* in some world *w*," "$\Box Fa$" translates as "*F* holds of every counterpart *c* of *a* in any world *w*," and so on. A sentence with no modal operator is translated into counterpart theory by restricting the domain of quantification to the actual world.[3]

For the purposes of this chapter, the most important element of Lewis's counterpart theory is the counterpart relation itself. And the crucial feature of the counterpart relation is that *it is not identity*. The basic idea is that individuals, such as Barack Obama, or the table in the Oval Office, are "world-bound": no individual exists in more than one possible world. However, individuals do have *counterparts* at other possible worlds. Many possible worlds have counterparts of Obama (and indeed his table) in them – individuals that resemble Obama sufficiently to count as counterparts of him. So for example, the sentence "Obama might never have become president" is true not because there is some possible world in which *Obama himself* never becomes president: Obama, being world-bound, only exists at the actual world, and so there is no such possible world. Rather, the sentence is true because there is some possible world in which a *counterpart* of Obama never becomes president. Similarly, "Fraser is essentially human" is true (if indeed it is true) not because Fraser *himself* is human at every world at which he exists; that claim is trivially true, given that Fraser is located in the actual world and is, actually, human. Rather, it is true because all of Fraser's other-worldly *counterparts* are human. According to counterpart theory, then, no individuals are "modal continuants." Neither Obama nor his desk is an object that is "spread out" across, or exists in, many different possible worlds; they are both firmly located in the actual world. Lewis thus denies the thesis of "transworld identity," about which we say a little more below.

What determines whether some object in some other possible world is or is not a counterpart of, say, Obama? Lewis's answer is: *similarity*. If there is an object *o* at some possible world *w* that is sufficiently similar to Obama, then *o* is Obama's counterpart at *w*. If there isn't, then Obama lacks a counterpart at *w*. As Lewis points out, the counterpart relation is therefore "problematic in the way all relations of similarity are: it is the resultant of similarities and dissimilarities in a multitude of respects, weighted by the importance of various respects and by the degrees of similarities" (1968, 115).[4]

How exactly does counterpart theory engage with Quine's worry about the senselessness of Aristotelian essentialism? Well, Lewis notes Quine's warning – and Quine's claim that "Aristotelian essentialism should be every bit as congenial to [the champion of quantified modal logic] as quantified modal logic itself" (Quine 1966, 182; quoted in Lewis 1968, 119). Lewis's tongue-in-cheek response is: "Agreed. Aristotelian essentialism is congenial" (119). He then proceeds to address Quine's doubts about the intelligibility of Aristotelian essentialism by defining essence in terms of the counterpart relation. Thus "*a* is essentially *F*" translates as "all of *a*'s counterparts are *F*," and "*a* is accidentally *F*" translates as "*a* is *F*, but not all of *a*'s counterparts are *F*."

Does counterpart theory really make *sense* of Aristotelian essentialism? Well, Lewis says:

> Whereas the obscurity of quantified modal logic has proved intractable, that of counterpart theory is at least divided, if not conquered. We can trace it to its two independent sources. There is our uncertainty about analyticity, and, hence, about whether certain descriptions describe possible worlds; and there is our uncertainty about the relative importance of different respects of similarity and dissimilarity, and, hence, about which things are counterparts of which. (1968, 116–17)

We take this to be an indication that, according to Lewis, counterpart theory at least makes *some* progress on the issue of Aristotelian essentialism. Quine could not see any glimmer of a way of determining what the essential and accidental features of an object, considered in and of itself, might be, and counterpart theory provides – at least in principle – a method for doing so. For our way of figuring out whether, say, Gavin is essentially rational is to figure out whether all of Gavin's counterparts are rational; and we do *that* by figuring out whether all of the inhabitants of other possible worlds that are similar enough to him to *count* as his counterparts are rational. And while all this may yet prove extremely difficult to figure out, Lewis does appear to think that the problem is tractable; after all, if he did not think this, he would have made very little progress at all in answering Quine's concern. Note that whilst he here says that we are "uncertain" about the relative importance of different respects of similarity, he appears to regard disputes about "which things are counterparts of which" as resolvable at least in principle.

Note that a commitment to transworld identity, rather than the counterpart relation, would not deliver any progress when it comes to Quine's worry. If our question about whether or not Gavin is essentially rational is to be answered by whether or not some possible irrational individual *is* Gavin, we are no further forward. In particular we cannot answer the question whether a given possible individual *is* Gavin by appealing to similarity because similarity is neither necessary nor sufficient for identity (see Kripke 1981, 43).

Further textual evidence that Lewis (1968) does indeed take his appeal to similarity to address Quine's bewilderment comes from his reference (1968, 115, n.4) to Michael Slote's "The Theory of Important Criteria" (1966). The gist of Slote's argument is that the meanings of many general terms are determined by the "Principle of Important Criteria" (PIC): "*x* is (an) *f* if and only if *x* has all the important criteria of *f*-ness" (1966, 215). So disputes about, say, whether a whale is a fish or a mammal arise from a difference of opinion between biologists and those ignorant of biology of what the "important" criteria for fishhood or mammalhood are, with the ignorant folk thinking that scales and living in water are important and biologists favoring other criteria. Similarly, a dispute about whether Buffalo is a city might arise because one party thinks that population size is the only important criterion, while another thinks that "the possession of a rich and varied cultural life and the possession of a sophisticated, cosmopolitan, urbane atmosphere" are also important (1966, 217). In the case of the whale, Slote thinks, the biologists are right: the biologists' criteria, which classify the whale as a mammal, really are more important than those of the ordinary person, since they

> are important for disinterested knowledge about and understanding of animals, that to learn whether some animal *x* has or lacks one of these characteristics is to learn something important for our disinterested understanding of *x*, i.e., for our knowledge of the sort of thing *x* is. (1966, 219)

In the case of Buffalo the issue turns on matters of taste, and may therefore not be resolvable.

While the appeal to Slote may provide some textual evidence that Lewis, circa 1968, took "our uncertainty about the relative importance of different respects of similarity and dissimilarity" – and hence our uncertainty about which features of an object are essential and which accidental – to be

resolvable in principle, it is unclear why he would think that Slote's view really gets us anywhere when it comes to counterpart relations (and hence essences). For Slote's view is a view about *general* terms, while the counterpart relation is a relation between individuals. Even supposing that we grant that the dispute about whales is resolved, so that, necessarily, nothing that is a whale is a fish, this does not, just by itself, help us decide whether any *particular* whale – Moby Dick, say – has any fish as counterparts, and hence whether being a mammal is part of Moby Dick's essence, since Slote's principle is silent on whether anything that is a whale is necessarily a whale. This is a question primarily for the metaphysician, and one that Lewis himself shows no inclination towards addressing here.

Considered as a response to Quine's concern about the intelligibility of Aristotelian essentialism, then, counterpart theory at best only gets us part of the way. We know what determines facts about essences according to Lewis, namely facts about the similarity relations that obtain between possible individuals. But we are no closer to actually addressing Quine's bewilderment about the cycling mathematician. Indeed, Lewis's appeal to Slote looks suspiciously like sidestepping that question. Quine is (relatively) happy to grant that there are necessary truths such as "all mathematicians are rational" and (to use Slote's example) "all whales are mammals." But these are *de dicto* rather than *de re* modal truths, and hence tell us nothing about the essence of any particular whale, any more than they tell us anything about the essence of any particular cycling mathematician.

14.3 Counterpart Theory and the Inconstancy of *De Re* Modal Predication

Here's a slightly different way of summing up the point we've reached. Quine was skeptical about Aristotelian essentialism because he could see no means of drawing a distinction that isn't invidious between the essential and accidental attributes of a thing. Lewis's initial response to Quine's skepticism was to say that actually there is such a distinction to be drawn, namely, between the attributes of a thing x that are shared by all its counterparts (aka the essential attributes of x) and the attributes that aren't (aka the accidental ones). But unless we are already equipped with a distinction that isn't invidious between the things that are counterparts of x and the things that aren't, Quine's skepticism will be nowise answered. Lewis's initial development of counterpart theory assumes that appeal to similarity would provide the basis for such a distinction. But since Lewis's original version of counterpart theory fails to provide a principled means for weighting the relative importance of different respects of similarity when it comes to individuals and their counterparts, it can hardly be deemed a satisfactory response to Quine's skepticism.

In his original formulation of counterpart theory (1968) Lewis assumes that the counterpart relation is unique. But he was soon to jettison this assumption. His "Counterparts of Persons and Their Bodies" (1971) draws explicit attention to this change of heart: "it seems necessary to revise my counterpart theory by providing for a multiplicity of counterpart relations" (1971, 203). Lewis's immediate concern in this paper is to address puzzles about personal identity. But his recognition of a multiplicity of counterpart relations also tells us something about the character of the far deeper response to Quine's skepticism that runs through Lewis's subsequent philosophy. It tells us that Lewis agrees with Quine that we have no determinate counterpart relation; so to this extent Lewis agrees with Quine's misgivings about essentialism, that is, insofar as the target notion of essence presupposes the availability of such a unique relation.

It will help us get clearer about the subtle relationship between Quine's skepticism and Lewis's counterpart theory if we expose ourselves once more to Quine's efforts to evoke bewilderment about the difference between essential and accidental attributes. About the concrete individual, Gavin, who (we like to imagine) proves whilst he pedals, Quine writes,

> Just insofar as we are talking referentially of the object, with no special bias toward a background grouping of mathematicians as against cyclists or vice versa, there is no semblance of sense in rating some of his attributes as necessary and others as contingent. (1960, 199)

Now it is indeed the case that Quine is making in this passage the negative (skeptical) claim that when we disregard a "special bias towards a background grouping" there is no sensible answer we can give to the question, "Is this concrete individual necessarily rational and contingently two-legged or vice versa?" But what is less often noticed is that Quine is also making a positive suggestion in this passage, namely, that when we do pay regard to a background grouping there *is* a semblance of sense in rating some attributes as necessary, others contingent. If the focus of our inquiry is mathematicians then the question whether they are essentially rational will receive one answer ("yes"), whereas if the focus of interest is cyclists another answer will be appropriate ("no"). Quine elaborates upon this suggestion in "Intensions Revisited" (1977), where he maintained that the notion of essence makes "sense in context":

> Relative to a particular inquiry, some predicates may play a more basic role than others, or may apply more fixedly; and these may be treated as essential. The respective derivative notions, then, of vivid designator and rigid designator, are similarly dependent on context and empty otherwise. (1977, 121)

How does this square with Quine's die-hard skepticism about essentialism? What these passages show us is that Quine only rejects the idea of *de re* necessity insofar as it is understood to rely upon a notion of essence that isn't context-relative. It is against this absolute conception of *de re* necessity that Quine inveighs when he denies the intelligibility of the question, which traits of a concrete individual are essential, which accidental. Of course, Quine also deemed contextually sensitive idioms that were "sources of truth-value fluctuation" ultimately dispensable for scientific purposes, that is, for an "austere canonical form for the system of the world" (Quine 1960, 228). But let's keep this last point in abeyance. What's important for now is that, so far from disagreeing with Quine about the intelligibility of *de re* necessity conceived independently of context, Lewis concurs entirely with him. In fact Lewis picked up the ball and ran with Quine's idea that *de re* modal predications only make "sense in context"; his developed counterpart theory is the result.

We have already taken note of one key revision that separates Lewis's original version of counterpart theory from his subsequent elaboration upon it: the recognition of a multiplicity of counterpart relations. According to Lewis, "the" counterpart relation is a relation of comparative overall similarity and comparative overall similarity is a vague or indeterminate notion (Lewis 1971, 211; 1983a, 42). The notion is vague because there is no absolute standard that settles whether certain respects of similarity and difference count or not, whether one respect counts more than another, what minimal standard of similarity is required, and so on: "as with vagueness generally, the vagueness of the counterpart relation – and hence of essence and *de re* modality generally – may be subject to pragmatic pressures, and differently resolved in different contexts" (1983a, 42). In one context we may favor the counterpart relation that (e.g.) weighs being two-legged more heavily than being rational, but in another context we may reverse the relative weighting of these respects of similarity, and so on. Thinking along these lines we recognize an extraordinary multiplicity of counterpart relations corresponding to all the different possible resolutions of the vagueness inherent in the notion of comparative overall similarity. It is only context that makes one or some of these relations salient to our assessment of (e.g.) whether our mathematician-cyclist is essentially two-legged. Once a context *C* has enabled us to fix upon a sufficiently determinate counterpart relation *R*, we are then in a position to raise the significant question of whether all and only the individuals that are *R*-related to him are two legged; if so, then, relative to *C*, he is essentially two-legged. But absent the pressure

of context to resolve the indeterminacy of the counterpart relation, the question of whether some or other attributes of a thing are essential or accidental to it, isn't really asking anything. Without the determining strictures of context the language of essence just goes on holiday. Lewis brings his case home by inviting us to answer the following questions: "Could Hubert Humphrey have been an angel? A human born to different parents? A human born to different parents in ancient Egypt? A robot? A clever donkey that talks? An ordinary donkey? A poached egg?" (1986a, 251). Whilst Lewis thinks that given "some contextual guidance these questions should have sensible answers," he also predicts that our immediate response will be one of bewilderment. Why so? He explains: "Your problem is that the right way of representing him is determined, or perhaps underdetermined, by context – and I supplied no context" (251).

Why does Lewis think we are compelled to think along these lines? Wittgenstein famously argued against essentialism on the grounds that when we pay attention to the complicated network of different ways in which we use (e.g.) the term "game," we will see only a "family resemblance" between the things we pick out when we use this term (Wittgenstein 1953, §§65–6). Lewis was no less impressed by what he described as the "inconstancy" of the language of essence. In *On the Plurality of Worlds*, Lewis asks us to "[a]ttend to the variety of what we say about modality and counterfactuals *de re*" (1986a, 252). What he thought we would find there, if we only took care to inspect, is "abundant evidence that we do not have settled answers, fixed once and for all" about which attributes of a thing are essential, which accidental: "different answers are often right in different contexts." With the benefit of hindsight we may interpret Quine's skepticism about modality as flowing from an inchoate appreciation of the inconstancy of modal discourse that Lewis draws to our attention. Quine's instinct was to recommend that we dispense with such unruly discourse for scientific purposes. But Lewis's response was to insist upon the respectability of ordinary discourse. What he provides – and what Quine only hinted at – is a way of understanding how it is possible for inconstant discourse to be rule-governed. It can be rule-governed, even though inconstant, because the fluctuations in the truth-values of modal constructions are merely a symptom of our resolving the indeterminacy of comparative overall similarity in different ways in different contexts. Hence: "Inconstancy in representation *de re* is exactly what we should expect under the hypothesis that it works by comparative overall similarity of complex things" (1986a, 254). The inconstancy of representation *de re* accordingly counts in favor of Lewis's thesis that it is only relative to a context that the distinction between the essential and accidental attributes of a thing makes any sense.

We have so far emphasized the extent to which Lewis's counterpart theory emerges out of Quine's skepticism about *de re* modality. But it will help bring the different aspects of our discussion together if we focus on an important difference between them. Recall once more Quine's original attempt to evoke bewilderment about Aristotelian essentialism. He pointed that we are unable to determine once and for all whether an attribute of our mathematician-cyclist is essential or accidental without privileging one description of him at the expense of the other. Quine took this fact to betoken our failure to refer *de re* to the concrete individual in question when we distinguish between the essential and accidental attributes it bears. But Lewis took the fact differently: the context-dependence of essence and accident no more betokens a failure of representation *de re* than does the fact that we cannot settle whether Tian Tian is small without having a description of her as (e.g.) a giant panda rather than as a mammal in mind. Obviously this doesn't show that we fail to refer *de re* to Tian Tian when we say that she's small: *she* is small relative to the class of giant pandas even though *she* isn't small relative to the class of mammals.

Of course the Aristotelian essentialist against whom Quine originally inveighed is unlikely to find Lewis's relativism satisfactory. The suggestion that there is no absolute answer to the question which of a thing's attributes are essential, which accidental, is likely only to induce a sense of metaphysical

vertigo in such a character. L.A. Paul expresses this sense of felt dissatisfaction with Lewis's counterpart theory when she reflects:

> What objects *are* is determined by their modal properties, and these are determined in a context independent manner…If the essential and accidental properties of an object are not absolute, then there are no absolute facts about what an object has to be like in order to exist. And this seems wrong. (2006, 345)

Lewis anticipates this response to the use he has made of counterpart theory: "The true-hearted essentialist might well think me a false friend, a Quinean sceptic in essentialist's clothing" (1983a, 42). We are now in a position to appreciate why Lewis is unmoved by the accusation. In order to understand how there could be an absolute fact of the matter about which of a thing's attributes are essential, which accidental, we should have to conceive of the counterpart relation itself as completely determinate, thereby obviating the need to rely upon context to resolve the indeterminacy. But what the inconstancy of modal language forces upon us is the recognition that the counterpart relation *is* indeterminate – otherwise the inconstancy of our language would betoken nothing but its failure to be rule-governed. And once we have recognized this we appreciate that it makes no more sense to ask whether something has an attribute essentially or accidentally independently of context than it makes sense to ask whether something is large or small without having a comparison class in mind.

14.4 Truthmaking and Counterpart Theory

In the preceding section we showed how Lewis's counterpart theory allows him to acknowledge the force of Quine's skepticism about Aristotelian essentialism whilst also enabling him to make sense of QML in particular, and *de re* modal language more generally. Indeed one might think of Lewis as offering a "skeptical solution" to the problems that Quine poses for *de re* modality: there are no absolute facts about the modal profile of a thing. There had better not be, otherwise we would be unable to accommodate the inconstancy of modal discourse; nonetheless, the pressure of context allows us to discriminate, without being invidious, between the attributes of a thing which are essential and those which are accidental. This already suggests one sense in which Lewis may be cast as a Humean – compare Hume's "skeptical solution" to the problem of induction. Of course Lewis is well known for his commitment to the doctrine of Humean supervenience: "the doctrine that all there is to the world is a vast mosaic of local matters of particular fact, just one little thing and then another" (1986b: ix). But, we will argue in the rest of this chapter, Lewis's world-view is far more thoroughly Humean than his commitment to this doctrine suggests. We base our case upon Lewis's use of counterpart theory to serve a much broader and more radical agenda, to eschew necessary connections between distinct existences and to avoid the idle disputes of "school metaphysics."

Recognizing the inconstancy of modal discourse and appreciating its true source – the vagueness of the counterpart relation, not only enables us to validate what we ordinarily say about modality and counterfactuals *de re*. If Lewis is right, doing so also enables us to dissolve a variety of metaphysical puzzles. Consider a statue, Goliath, and the lump of clay, Lumpl, from which it is composed (Gibbard 1975). The lump remains in that shape until it is destroyed. So Goliath and Lumpl entirely coincide in their spatiotemporal extent. This makes it tempting to conclude that they are identical, that Goliath *is* Lumpl. But Lumpl could have survived squashing whereas Goliath could not; the attributes that are essential to Goliath are only accidental to Lumpl. How could this be the case if Lumpl and Goliath are one and the same? It appears that we are committed to an unwanted

multiplication of entities, distinct things that coincide at all times of their existence, distinguished only by their modal and counterfactual attributes. According to Lewis, the appearance of such an unwanted multiplication of entities is generated only by a failure to factor in the inconstancy of modal discourse (1971). When we say (i) that Goliath is essentially the shape of a statue of Goliath, our use of the name "Goliath" evokes the counterpart relation for statues. Whereas when we say (ii) that Lumpl isn't essentially but only accidentally so, our use of the name "Lumpl" evokes a different counterpart relation, one for lumps. Because they evoke different counterpart relations, what (i) affirms of Goliath isn't what (ii) denies of Lumpl. It follows there is no incompatibility generated by supposing that one and the same subject is represented *de re* by both (i) and (ii). So there is no need to distinguish Goliath from Lumpl after all.

Appeal to the multiplicity of counterpart relations thereby provides Lewis with a mechanism for avoiding the multiplication of entities. It also enables him to avoid the need to explain what he describes as the "strangely intimate relationship of constitution" that would have obtained between Goliath and Lumpl (and other lumps and statues) if they had been distinguished. It is to serve this latter role, the avoidance of such connections, that Lewis deploys counterpart theory in "Things *qua* Truthmakers" (2003).

Lewis was a long-term critic of the idea that a proposition *P* is made true by a thing *T* whose mere existence necessitates the truth of *P*, and the Truthmaker Principle that says every truth has a truthmaker in this sense (1992, 1998, 2001). His main complaint was that a commitment to such truthmakers generates a wake of unwanted necessary connections between wholly distinct existences. Lewis accordingly rejected the demand for truthmakers "[b]ecause, sadly, the demand for truthmakers just *is* a demand for necessary connections" (1998, 219).

One way to appreciate how the unwanted connections are thrown up by the churning of the Truthmaker Principle is to focus on the fact that, with respect to many of the true things we say, ordinary objects are incapable of discharging the role of making those things true. Suppose an object *A* has an intrinsic property *F*, but only accidentally: *A* could have lacked the property *F*. Then *A* cannot make it true that *A* has *F*, because it has just been granted that the mere existence of *A* fails to necessitate its being *F*. Nor can *A* and *F* together make it true that *A* is *F* because *A* could have existed whilst something else was *F* instead. The only entity that has the relevant truthmaking potential is the *state of affairs* that *A* is *F* – because this is an entity that cannot exist unless *A* is *F*. But Lewis refuses to admit states of affairs. His rejection of them relies upon his finding "unmereological 'composition' profoundly mysterious"; so far as Lewis was concerned mereology is the only kind of composition (1991, 57; see also 1986c, 39; 1986d). Now the state of affairs that *A* is *F*, Δ, is not composed mereologically from its "constituents" *A* and *F*. So it follows from Lewis's assumption that mereology is the only kind of composition that Δ isn't composed from *A* and *F* but is entirely distinct from them. But even though Δ is entirely distinct from *A* and *F*, it is necessarily connected to them: necessarily if Δ exists then *A* does, *F* does, and so on (Lewis 1998, 49). Subsequently Lewis was to admit that he could understand unmereological composition after all – but only by defining it in terms of necessary connections! "To say that the state of affairs is unmereologically composed of *F* and *A* is to say nothing more or less than that, necessarily, it exists iff *A* has *F*" (2001, 113). But he still complained that by describing the composition of Δ as unmereological "a necessary connection between seemingly distinct existences has been in no way explained, or excused" (113). Since his Humean suspicions of necessary connections have been nowise appeased, Lewis refuses to acknowledge states of affairs and rejects the Truthmaker Principle that requires them (113).

Lewis is especially suspicious of the necessary connections thrown up when we posit truthmakers for negative existentials. He declares such posits "bad news for systematic metaphysics" (2001, 111). Why so? Suppose there is a truth maker Γ for the negative existential truth *P*: there are no unicorns. Since Γ is a truth maker for *P*, it cannot coexist with any unicorn. Γ, so to speak, is a

unicorn-replacement. It is the capacity of Γ to exclude the existence of things entirely distinct from it, namely unicorns, that makes it eligible to be a truthmaker for *P*. But it is exactly this feature of Γ that makes it unacceptable from Lewis's point of view:

> in order to do its job as a truthmaker, the unicorn-replacement must be involved in necessary connections between (mereologically) distinct existences; and it is the Humean prohibition of necessary connections that gives our best handle on the question what possibilities there are. (2001, 111)

Can the mystery surrounding these necessary connections be cleared up? Lewis came to recognize that if he availed himself of counterpart theory then he could adopt the Truthmaker Principle without compromising his eschewal of necessary connections. His key insight was the appreciation that to describe something as a truthmaker for a proposition *P* is to ascribe to it an essential attribute. It's important to keep clearly in mind the conception of truthmaking that Lewis adheres to when he says this. A truthmaker for *P* is something the mere existence of which suffices for the truth of *P*. So a truthmaker for *P* could not have existed without *P* being true. In other words, a truthmaker for *P* is essentially such that *P* is true.

Now the need to posit (e.g.) the state of affairs Δ as truthmaker for the claim that *A* is *F* was forced upon us because of what seemed an innocent assumption, namely, that *A* doesn't have the essential attribute of being *F*. After all, how could *A* conceivably have this essential attribute when it is only contingently true that *A* is *F*? We concluded that *A* couldn't be a truthmaker for the claim that *A* is *F*. But once we have taken on board the inconstancy of *de re* modal discourse, this assumption no longer seems so innocent. There is no absolute fact of the matter about whether *A* is essentially such that *A* is *F*; it depends upon which counterpart relation is evoked when the claim is made. If it is one that tolerates counterparts of *A* that fail to be *F* then the claim is false, whereas if it is a more demanding relation that includes only *F* things amongst the counterparts of *A* then the claim is true.

Lewis's strategy for avoiding states of affairs – and their concomitant necessary connections – whilst adhering to the Truthmaker Principle was to introduce a novel "qua" locution that can be used to evoke the more demanding kind of counterpart relation relative to which *A* is essentially such that *A* is *F* (Lewis 2003, 30–1). For example, the phrase "Tian Tian qua hungry" evokes the counterpart relation whereby every one of Tian Tian's counterparts is guaranteed to be hungry. So Tian Tian qua hungry does have the essential attribute of being such that she is hungry (because every one of her counterparts is). Hence Tian Tian qua hungry is a truthmaker for "Tian Tan is hungry." Does this mean that Tian Tian qua hungry is a novel and peculiar kind of thing, perhaps even a state of affairs in disguise? Not at all. Whilst Tian Tian qua hungry makes it true that she's hungry, Tian Tian doesn't – it's only one of her accidental attributes that she's hungry. But it doesn't follow that Tian Tian qua hungry isn't Tian Tian. Because the names evoke different counterpart relations, what we affirm when we say (iii) that Tian Tian qua hungry makes it true that Tian Tian is hungry isn't what we deny when we say (iv) that Tian Tian doesn't. It follows that there's no incompatibility in supposing that it is one and the same subject represented *de re* by (iii) and (iv).

We might extend Lewis's account at this point by adding the speculative suggestion that the "state-of-affairs" idiom is simply a more familiar version of Lewis's "qua" construction.[5] Then even though the state of affairs that *A* is *F* is a truthmaker for "*A* is *F*," whilst *A* isn't, it doesn't follow that the states of affairs that *A* is *F* isn't *A*. In this way the appeal to the multiplicity of counterpart relations enables us to avoid a multitude of states of affairs, and obviates the need to understand the "necessary connections" that obtains between states of affairs and their constituents, just as it enabled us to avoid distinguishing Goliath from Lumpl and the "strangely intimate relationship" that supposedly unites them.

A further application of counterpart theory supplies truthmakers for negative existentials too (Lewis and Rosen 2003). Suppose, to take the simplest case, that there are no *F*s whatsoever. Then *F*s are entirely absent from the world. Does the world then make it true there are no *F*s? Hardly. The world has counterparts to which *F*s have been added. But take the world qua unaccompanied by *F*s. It essentially lacks *F*s. So it is a truthmaker for the claim that there are no *F*s. But this doesn't embroil us in necessary connections because the world qua unaccompanied by unicorns is just the world by another name.

14.5 Lewis versus School Metaphysics

In the famous closing paragraph of the first *Enquiry*, Hume says:

> When we run over libraries, persuaded of these principles, what havoc must we make? If we take in our hand any volume of divinity or school metaphysics, for instance, let us ask, *Does it contain any abstract reasoning concerning quantity or number?* No. *Does it contain any experimental reasoning concerning matter of fact and existence?* No. Commit it then to the flames, for it can contain nothing but sophistry and illusion. (1975, §12.3, 165)

Lewis does not of course sign up to Hume's account of the legitimate sources of human knowledge; and of course an interest in the failings of "school metaphysics" – namely Aristotelian scholasticism – would be somewhat anachronistic. Nonetheless, we shall argue in this section that Lewis's attitude to some debates in contemporary ontology, again deriving from the inconstancy of modality, is on a par with Hume's attitude to school metaphysics – and for broadly analogous reasons.

In *On the Plurality of Worlds* and elsewhere, Lewis provides some examples of cases where shifts in context prompt the appearance of a genuine ontological dispute about the *de re* modal facts when in fact, on his view, there is no such dispute to be had. One is the dispute about the essentiality of origin:

> I suggest that those philosophers who preach that origins are essential are absolutely right – in the context of their own preaching. They make themselves right: their preaching constitutes a context in which *de re* modality is governed by a way of representing (as I think, by a counterpart relation) that requires match of origins. But if I ask how things would be if Saul Kripke had come from no sperm and egg but had been brought by a stork, that makes equally good sense. I create a context that makes my question makes sense, and to do so it has to be a context that makes origins not be essential. (1986a, 252)

The disputes about the relation between a thing and the matter that constitutes it (1986a, 257) and about mereological essentialism (1986a, 248–51) get similar treatments.

These are all standard disputes in contemporary ontology, and none of them, in Lewis's book, constitutes a genuine dispute at all. As he puts it:

> You could do worse than plunge for the first answer to come into your head, and defend that strenuously. If you did, your answer would be right. For your answer itself would create a context, and the context would select a way of representing, and the way of representing would be such as to make your answer true. (1986a, 251)

Moreover, it is not merely Lewis's implied judgment about the vacuity of such disputes about *de re* modality that aligns him with Hume (and we might speculate that Lewis's use of the word "preach" in the above passage is a nod towards Hume's antipathy towards scholasticism). The presupposition

that ontological "disputes" such as these are bona fide disputes brings in its train the view that our ordinary *de re* modal judgments place epistemic constraints on our metaphysical theorizing. We shall argue that from Lewis's perspective – according to which the relevant constraints are methodological rather than epistemic – such disputes engender a decidedly scholastic conception of our cognitive access to the fundamental nature of reality.

As we have seen, Lewis's commitment to the context-dependence of the counterpart relation derives from an appreciation of the inconstancy of our ordinary, common sense *de re* modal judgments. For Lewis, common sense is, as far as possible, not to be violated; and so (as he says in a variety of contexts), where our ordinary judgments are not settled, it is a virtue of our philosophical theory if it doesn't deliver a decisive answer either. Now, why does Lewis take this attitude to common sense judgments, and what does he mean by "common sense"?

Lewis's answer to both questions comes in the following passage:

> . . . a theory cannot earn credence just by its unity and economy. What credence it cannot earn, it must inherit. It is far beyond our power to weave a brand new fabric of adequate theory *ex nihilo*, so we must perforce conserve the one we've got. A worthwhile theory must be credible, and a credible theory must be conservative. It cannot gain, and it cannot deserve, credence if it disagrees with too much of common sense. Common sense is a settled body of theory – unsystematic folk theory – which at any rate we *do* believe; and I presume that we are reasonable to believe it. (*Most* of it.) (1986a, 134)

Our philosophical theorizing is thus, for Lewis, constrained by *our* common sense theory of the world: there are not many different "common sense" theories held by different people. Thus, for example, where two people disagree with each other, say about whether Kripke could have been brought by a stork, for Lewis that constitutes evidence for the inconstancy of *our* – collective – common sense theory even if the participants in the dispute individually exhibit no inconstancy at all. (Perhaps one participant has had their contextual parameters fixed once and for all by early exposure to *Naming and Necessity*, and the other by their ability to make perfectly good sense of the fairy story they were told as a child concerning the provenance of babies.) And our common sense theory – and therewith the inconstancy of *de re* modal judgments – is to be respected as far as possible simply because philosophical theory construction demands conservatism: we lack the intellectual resources to start from scratch. Thus, for Lewis, the truth (or close to it) of our common sense theory imposes a methodological constraint on our theorizing: we can undertake minor repairs to Neurath's boat if necessary, but since the boat is seaworthy and indeed already out on the high seas, large-scale repairs would be both unnecessary and dangerous.

This view of common sense theory contrasts sharply with the view taken in much contemporary debate about *de re* modality. Suppose we start out with the assumption that a dispute about, say, the essentiality of origin is a genuine dispute, so that there is an absolute fact of the matter about whether Kripke could have been brought by a stork. We cannot now regard the "dispute" about whether or not Kripke could have been brought by a stork as merely an instance of the inconstancy of our collective common sense theory, where inconstancy itself is something that our metaphysical theory should respect; that option has been foreclosed by the presupposition that there is a metaphysical fact of the matter that our dispute is aimed at resolving. Hence *our* common sense theory, as Lewis conceives it, cannot help us.

How, then, *is* the dispute to be resolved? Well, it is a dispute about the nature of reality, and so our method for resolving it must, somehow or other, make contact with reality. And yet we cannot invoke experience or observation, since we're doing metaphysics, not science; so where should we turn? The standard answer is: "intuition." The nature and epistemic status of "metaphysical intuition" is itself currently much in dispute; for example Paul describes intuitions as "quasi-perceptual" (2010, 462), while according to Ernest Sosa: "When we rely on intuitions in philosophy . . . we manifest a

competence that enables us to get it right on a certain subject matter, by basing our beliefs on the sheer understanding of their contents" (2007, 102). But what is generally agreed upon is that metaphysical intuitions have epistemic status: that a particular philosopher *X* has the intuition that Kripke could have been brought by a stork is, by *X*'s lights, at least prima facie reason to think that the essentiality of origin is false.

If we think of our own judgments about Kripke's origins as equivalent to or arising out of intuitions to which at least prima facie epistemic entitlement attaches, then of course inconstancy, as exhibited by the firmly held but contradictory views of the disputants, is likely to seem anomalous and unsettling: analogous to two experimenters peering through the same microscope and claiming to see entirely different things. In the case of our experimenters, we quite plausibly assume that something has gone awry: maybe one of them needs a new pair of spectacles or someone twiddled the dial between observations. And so, on this way of viewing intuitions, we are liable to see inconstancy as a kind of cognitive failing on the part of one of the disputants: there is a fact of the matter to be had, but one of us just isn't *getting* it.

We are now in a position to explain our earlier claim that, from Lewis's point of view, the supposition that "disputes" about *de re* modality are genuine disputes engenders a decidedly scholastic – and from Lewis's point of view unwelcome – conception of our cognitive access to the fundamental nature of reality. First, we can only suppose that there are genuine disputes to be had here if we jettison Lewis's conception of the nature and methodological significance of "common sense theory," since on Lewis's conception what common sense reveals is that our *de re* modal judgments are true or false not absolutely but relative to a context. And, second, the replacement for (our collective) common sense theory as a methodological constraint on metaphysical theorizing is a view according to which our "intuitions" about the absolute *de re* modal features of reality are to be accorded epistemic status. And this really is a move towards scholasticism. Hume held that all our knowledge of facts about the external world comes to us through our senses or through reasoning from causes to effects, and the scholastics, who held that we could penetrate into the essences of things, were amongst his prime targets. But this – for Lewis no less than for Hume – cannot be done.

14.6 Humean Supervenience and *De Re* Modality

In sections 14.4 and 14.5, we explored two ways in which Lewis's views on *de re* modality should be seen as distinctly Humean. But we have so far placed little weight upon Lewis's well-known commitment to the doctrine of Humean supervenience: "the doctrine that all there is to the world is a vast mosaic of local matters of particular fact, just one little thing and then another" (1986a, ix). In this section, we explain how Lewis's eschewal of necessary connections between distinct existences – already discussed in the context of truthmaker theory in section 14.4 above – connects with the thesis of Humean supervenience, and we argue that his world-view is far more thoroughly Humean than his commitment to the contingent truth of this doctrine suggests.

Counterpart theory is used by Lewis to provide a reductive analysis of necessity and possibility. Claims involving "necessarily" and "possibly" – and also "essentially," "contingently," and "accidentally" – can systematically be translated into claims about possible worlds and their inhabitants, and counterpart relations between those inhabitants, where counterpart relations themselves are similarity relations. To this account, Lewis later adds an account of the semantics of counterfactuals – which, again, appeals only to possible worlds and counterpart relations. And Lewis's concerns here are clearly not merely logical and semantic; his modal ontology is, precisely, an ontology of possible worlds – conceived as distinct concrete particulars – and similarity relations.

But what about the constituents of possible worlds (including the actual world) themselves? As is very well known, Lewis upholds the thesis of Humean supervenience:

> Humean Supervenience is named in honor of the greater denier of necessary connections. It is the doctrine that all there is to the world is a vast mosaic of local matters of particular fact, just one little thing and then another . . . We have geometry: a system of external relations of spatiotemporal distances between point . . . And at those points we have local qualities: perfectly natural intrinsic properties which need nothing bigger than a point at which to be instantiated. For short: we have an arrangement of qualities. And that is all. There is no difference without difference in the arrangement of qualities. All else supervenes on that. (1986b, ix)

Note that Lewis describes Hume as the *greater* denier of necessary connections, and not – contrary to popular misquotation – the *great* denier of necessary connections. Lewis is here clearly advertising himself as a denier of necessary connections. And so he is, at least when it comes to the actual world: neither Lewis's "qualities" nor the "arrangement" thereof have any modal features whatsoever, in and of themselves. Lewis's "qualities" are instances of perfectly natural properties, which are all categorical (see Lewis 2009), and their arrangement is purely spatiotemporal. Of course, there are modal truths about the qualities and their arrangement: two instantiations of the same perfectly natural property necessarily resemble each other, since natural properties just *are* those properties "whose sharing makes for resemblance" (Lewis 1983b, 13); the triangle formed by three equidistant points will necessarily be congruent with a triangle formed by any distinct trio of equidistant points; and so on. But those modal truths, like any other, are to be analyzed in terms of possible worlds and counterparts: they are not inherent in the very fabric of the Humean supervenience base itself.

Above the Humean supervenience base Lewis allows for modal connections between distinct existences, but they are always analyzable out by deploying the machinery of counterpart theory. Causes and effects, for example, are distinct existences that stand in a modal relation to each other. But causation itself is analyzed in terms of counterfactual dependence, which in turn gets an analysis in terms of similarity relations between possible worlds (Lewis 1973b, 1979). Or we might define a notion of physical necessitation, so that event x physically necessitates event y if and only if the occurrence of y is entailed by the occurrence of x together with the laws of nature. So we have – explicitly – necessary connections between distinct existences, but again the necessity is analysed in terms of similarity between worlds (since that is how Lewis analyzes lawhood – about which more below) together with straightforward entailment.

Of course, Humean supervenience is, by Lewis's own admission, at best contingent (at best because it may yet prove to be false). As he says:

> Two worlds might indeed differ only in unHumean ways, if one or both of them is a world where Humean supervenience fails. Perhaps there might be extra, irreducible external relations, besides the spatiotemporal ones . . . [etc.] It is not, alas, unintelligible that there might be suchlike rubbish. Some worlds have it. And when they do, it can make a difference between worlds even if they match perfectly in their arrangements of qualities.
>
> But if there is suchlike rubbish, say I, then there would have to be extra natural properties or relations that are altogether alien to this world. (1986b, x)

A world containing irreducible necessary connections between distinct existences would, of course, differ in "unHumean" ways from worlds that lack them. But could such necessary connections be "alien" relations, instantiated at other possible worlds but (it is to be hoped) not at the actual world? It would seem not. For – to repeat – modality is to be analyzed according to counterpart theory. Any modal truth whatsoever can be translated into a non-modal truth about similarity relations between

possible worlds. In other words, necessary connections fall outside the scope of Lewis's explanation of the contingency of Humean supervenience: necessary connections between distinct existences *necessarily* supervene on non-modal features. So there are no irreducible necessary connections between distinct existences at this or any other possible world.

Evidence for this claim – if such is needed – comes from Lewis's use of conceptual analysis. For example, while he is happy to admit the possibility of irreducibly mental states (one kind of "suchlike rubbish") – and his functionalist account of mental states leaves such a possibility open (Lewis 1966) – when it comes to *modal* features of the world, Lewis goes for full-blown reductionist conceptual analysis every time: necessary connections between distinct existences are to be "explained, or excused" (2001, 113). His accounts of laws (1973a, 72–7), causation (1973b), and dispositions (1997) in particular are all reductionist. The laws of nature are analyzed in terms of the arrangement of qualities; causal facts are analyzed in terms of facts about similarity of worlds, where non-violation of the laws constitutes one respect in which one world is similar to another; and the instantiation of dispositional properties is analyzed in terms of the instantiation of categorical properties plus (again) facts about the laws. Thus, for example, on Lewis's story the existence of fundamental dispositions is simply conceptually impossible, since what it *is* for something to have a dispositional property is for it to instantiate some categorical – and therefore non-modal – property such that its doing so, together with the laws, entails that it will behave in such-and-such a manner.

It is in the case of laws that Lewis comes closest to an explicit expression of a view that can be considered genuinely Humean. Armstrong argues that laws of nature are states of affairs in which a higher order relation of nomic necessitation (N) holds between lower order universals (F, G). According to Armstrong, the law that Fs are Gs consists in the obtaining of a higher-order state of affairs, $N(F,G)$. If the lower-order states of affairs Fa also obtains, then these two states of affairs together necessitate the further, distinct state of affairs Ga to obtain. But Armstrong's account of laws offends against Lewis's Humean sensibilities: "What leads me (with some regret) to reject Armstrong's theory . . . is that I find its necessary connections unintelligible. Whatever N may be, I cannot see how it could be absolutely impossible to have $N(F,G)$ and Fa without Ga" (1983b, 40). Of course $N(F, G)$ isn't *entirely* distinct from Ga but this reflection didn't appease Lewis; it only drew from him an expression of an even stronger suspicion about necessary connections:

> I am tempted to complain in Humean fashion of alleged necessary connections between distinct existences, especially when first-order states of affairs in the past supposedly join with second-order states of affairs to necessitate first-order states of affairs in the future. That complaint is not clearly right: the sharing of universals detracts from the distinctness of the necessitating and the necessitated states of affairs. But I am not appeased. I conclude that necessary connections can be unintelligible even when they are supposed to obtain between existences that are not clearly and wholly distinct. (1983b, 40)

Interpretation of Hume is, of course, a thorny business; but we think Lewis's position on necessary connections comes very close to Hume's remark that "we do not understand our own meaning" if we attempt to make the word "necessary connection" stand for some real, mind-independent feature of reality (1978, 1.3.14, 168) – with a caveat. For Lewis, it is not the reality or mind-independence of necessary connections between distinct existences that is being denied – causes and effects really do stand in such relations, for example, as a matter of mind-independent fact – but rather their status as constituents of this or any possible world, considered in isolation from the similarity relations possible worlds bear to each other.

On a standard and familiar reading of Hume, he reduces causation to constant conjunction. Lewis's metaphysics as a whole – and not merely his account of causation – can be seen as a more sophisticated and thoroughgoing application of Hume's (alleged) strategy for dealing with necessary connections. On a slightly less familiar reading of Hume (Beebee 2006, chapter 6; Coventry 2006),

causal facts are, in part, determined by the projection of our own minds onto a world that, in itself, contains nothing that could serve as the referent of the idea of necessary connection. On this way of looking at Hume, counterpart theory provides a Humean option that was not available to Hume himself: Hume was right that no referent for the idea of necessary connection is to be found in the *actual* world, but Lewis's much more generous conception of reality – it contains other possible worlds as well as our own – allows him to restore mind-independence to causation without giving up on Hume's central negative insight.

As we have said, it is in the case of laws that Lewis comes closest to explicitly aligning himself with an authentically Humean position. But the point we have been urging is that laws are not a special case for Lewis. On his conceptual analysis of laws, facts about what the laws are, at *any* possible world, wholly depend upon non-modal facts about the arrangement of qualities. But so, ultimately, do facts about dispositions and causation. Wherever a commitment to irreducible necessary connections between distinct existences threatens, Lewis responds with a conceptual analysis that removes the threat not just from the actual world, but from all possible worlds. And of course this is just what one would expect from a philosopher who regards such connections as unintelligible.

Notes

1 We henceforth drop the scare quotes from "Aristotelian essentialism," without intending to imply that what Quine means by the term has anything much to do with Aristotle.
2 Quine is not exactly sympathetic to *de dicto* modality either. He continues: "so much the worse for quantified modal logic. By implication, so much the worse for unquantified modal logic as well; for, if we do not propose to quantify across the necessity operator, the use of that operator ceases to have any clear advantage over merely quoting a sentence and saying that it is analytic" (1963, 156). Nor is he keen on analyticity itself (Quine 1951).
3 For some recent objections to counterpart theory, see Merricks 2003 and Fara and Williamson 2005.
4 There is a good deal of debate between Lewis and his opponents over the question of whether we should think of individuals as "worldbound" or "transworld" – that is, whether we should couch *de re* modality in terms of the counterpart relation (so that Obama is wholly located at the actual world) or in terms of transworld identity (so that the actual Obama is identical to various election-losing Obamas at other possible worlds) or in terms of mereology (so that the actual Obama is merely the actual-world part of the sum of all the Obamas, as it were, in something like the way his current self is merely a temporal part of the temporally extended object that came into existence in 1961). See, for example, Kripke 1980, 45; Plantinga 1973; and Lewis 1986a, chapter 4. However, as will become clear, it is the keen-eyed appreciation of the phenomenon of inconstancy that ultimately drives Lewis towards counterpart theory.
5 This extension of Lewis's deployment of counterpart theory to avoid necessary connections is envisaged in MacBride 2005, 139–40.

References

Armstrong, D.M. (1983). *What is a law of nature?* Cambridge: Cambridge University Press.
Beebee, H. (2006). *Hume on causation*. Abingdon: Routledge.
Coventry, A. (2006). *Hume's theory of causation: A quasi-realist interpretation*. London: Continuum.
Fara, M. & Williamson, T. (2005). Counterparts and actuality. *Mind*, 114, 1–30.
Gibbard, A. (1975). Contingent identity. *Journal of Philosophical Logic*, 4, 187–221.
Hume, D. (1978). *A treatise of human nature*. Ed. L.A. Selby-Bigge, 2nd ed. rev. and ed. P.H. Nidditch. Oxford: Clarendon Press. Orig. work 1739–40.

Hume, D. (1975). *Enquiries concerning human understanding and concerning the principles of morals*. Ed. L.A. Selby-Bigge, 3rd ed. rev. and ed. P.H. Nidditch. Oxford: Clarendon Press. Orig. work 1748–51.
Kripke, S. (1980). *Naming and necessity*. Oxford: Blackwell.
Lewis, D. (1966). An argument for the identity theory. *Journal of Philosophy*, 63, 17–25.
Lewis, D. (1968). Counterpart theory and quantified modal logic. *Journal of Philosophy*, 65, 113–26.
Lewis, D. (1971). Counterparts of persons and their bodies. *Journal of Philosophy*, 68, 203–11.
Lewis, D. (1973a). *Counterfactuals*. Oxford: Blackwell.
Lewis, D. (1973b). Causation. *Journal of Philosophy*, 70, 556–67.
Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, 13, 455–76.
Lewis, D. (1983a). Postscript to "Counterpart theory and quantified modal logic." In D. Lewis, *Philosophical papers*, Vol. I, pp. 39–46. New York: Oxford University Press.
Lewis, D. (1983b). New work for a theory of universals. *Australasian Journal of Philosophy*, 61, 343–77. Reprinted in D. Lewis, *Papers in metaphysics and epistemology* (1999), pp. 8–55. Cambridge: Cambridge University Press.
Lewis, D. (1986a). *On the plurality of worlds*. Oxford: Blackwell.
Lewis, D. (1986b). *Philosophical papers*, Vol. II. New York: Oxford University Press.
Lewis, D. (1986c). Against structural universals. *Australasian Journal of Philosophy*, 64, 25–46.
Lewis, D. (1986d). A comment on Armstrong and Forrest. *Australasian Journal of Philosophy*, 64, 92–3.
Lewis, D. (1991). *Parts of classes*. Oxford: Blackwell.
Lewis, D. (1992). Armstrong on combinatorial possibility. *Australasian Journal of Philosophy*, 64, 92–3.
Lewis, D. (1997). Finkish dispositions. *Philosophical Quarterly*, 47, 143–58.
Lewis, D. (1998). A world of truthmakers? *Times Literary Supplement*, 4950, 30.
Lewis, D. (2001). Truthmaking and difference-making. *Noûs*, 35, 602–15.
Lewis, D. (2003). Things *qua* truthmakers. In *Lillehammer & Rodriguez-Pereyra (2003)*, pp. 25–38.
Lewis, D. (2009). Ramseyan humility. In D. Braddon-Mitchell & R. Nola (Eds.), *Conceptual analysis and philosophical naturalism*, pp. 203–22. Cambridge, MA: MIT Press.
Lewis, D. & Rosen, G. (2003). Postscript to "Things *qua* truthmakers." In *Lillehammer & Rodriguez-Pereyra (2003)*, pp. 39–42.
Lillehammer, H. & Rodriguez-Pereyra, G. (Eds). (2003). *Real metaphysics: essays in honour of D.H. Mellor*. London: Routledge.
MacBride, F. (2005). Lewis's animadversions on the truthmaker principle. In H. Beebee & J. Dodd (Eds.), *Truthmakers: the contemporary debate*, pp. 117–40. Oxford: Oxford University Press.
Merricks, T. (2003). The end of counterpart theory. *Journal of Philosophy*, 100, 521–49.
Paul, L.A. (2006). In defense of essentialism. *Philosophical Perspectives*, 20, 333–72.
Paul, L.A. (2010). A new role for experimental work in metaphysics. *Review of Philosophy and Psychology*, 1, 461–76.
Plantinga, A. (1973). Transworld identity or worldbound individuals? In M. Munitz (Ed.), *Logic and ontology*, pp. 193–212. New York: New York University Press.
Quine, W.V. (1951). Two dogmas of empiricism. *Philosophical Review*, 60, 20–43.
Quine, W.V. (1960). *Word and object*. Cambridge, MA: MIT Press.
Quine, W.V. (1963). Reference and modality. In W.V. Quine, *From a logical point of view*, 2nd ed., pp. 139–59. New York: Harper & Row.
Quine, W.V. (1966). Reply to Professor Marcus. In W.V. Quine, *The ways of paradox*, p. 181. Cambridge, MA: Harvard University Press.
Quine, W.V. (1977). Intensions revisited. *Midwest Studies in Philosophy*, 2, 5–11.
Slote, M.A. (1966). The theory of important criteria. *Journal of Philosophy*, 63, 211–24.
Sosa, E. (2007). Experimental philosophy and philosophical intuition. *Philosophical Studies*, 132, 99–107.
Wittgenstein, L. (1953). *Philosophical investigations*. Trans. G.E.M. Anscombe. Oxford: Blackwell.

15

David Lewis on Persistence

KATHERINE HAWLEY

David Lewis takes a clear stance on persistence:

> Next, persistence through time. I take the view that nothing endures identically through time. (Except universals, if such there be; their loci would coincide with relations of qualitative match, would indeed constitute these relations, so they would commit no violations of Humean Supervenience.) Persisting particulars consist of temporal parts, united by various kinds of continuity. To the extent that the continuity is spatiotemporal and qualitative, of course it supervenes upon the arrangement of qualities. But the continuity that often matters most is causal continuity: the thing stays more or less the same because of the way its later temporal parts depend causally for their existence and character on the ones just before. So the spatiotemporal boundaries of persisting things, for example people, can supervene on the arrangement of qualities, provided that causation does. (Lewis 1986b, xiii)

To persist is to exist at more than one time, to transcend the momentary. How do things achieve this? We might answer with talk of thermodynamic stability, molecular bonds, photosynthesis, the porcupine's spines, German manufacturing standards, legal protection of ancient monuments, or the uncanny ability of children to extract care from their parents. In Lewis's terms, such answers explain the existence of spatiotemporal and qualitative continuities over time in causal terms, by reference either to the causal mechanisms which directly underpin such continuities, or to their preconditions and external circumstances. Explanations may differ according to the kind of object in question: German washing machines and yew trees are both long-lasting, relative to other types of appliance or tree respectively, but the reasons for their longevity are quite different.

The metaphysicians have a further question about persistence, a question which is expected to have the same answer for all sorts of concrete objects. What is it for something to exist at more than one time? For Lewis, a thing exists at more than one time by having distinct stages, each of which is located at a different time. These stages are parts of the persisting object: added together, they *are* the persisting object. On this view, persistence through time is analogous to extension through space: a spatially extended object occupies more than one point at a single time by having different spatial

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

parts located at different places. Likewise, for Lewis, a temporally extended object occupies more than one time by having different temporal parts located at different times. "Persisting particulars consist of temporal parts, united by various kinds of continuity."

The intended contrast is the view that concrete things "endure identically through time," as universals do if they exist. Universals are wholly located where their instances are: the universal *having mass 1 kg* is entirely present in each of the 1 kg bags of flour on the supermarket shelf, and that very same universal will be present in future bags of flour, just as it was in past bags of flour. The universal does not portion itself out, a part here and now, a part there and then: instead, the whole universal is in each place it is needed. In the same way, an enduring concrete thing, if such there be, would not portion itself out over time, a stage then and a stage now: instead, the whole persisting object is located at each time of its existence.

Following Lewis (who credits Mark Johnston), these rival views of persistence are now known as "perdurance theory" and "endurance theory" respectively:

> Let us say that something *persists* iff, somehow or other, it exists at various times; this is the neutral word. Something *perdures* iff it persists by having different temporal parts, or stages, at different times, though no one part of it is wholly present at more than one time; whereas it *endures* iff it persists by being wholly present at more than one time. Perdurance corresponds to the way a road persists through space; part of it is here and part of it is there, and no part of it is wholly present at two different places. Endurance corresponds to the way a universal, if there are such things, would be wholly present wherever and whenever it is instantiated. (Lewis 1986a, 202)

(Lewis had previously used "endure" as the neutral word ("Paradoxes of Time Travel," Lewis 1986b, 68).

Lewis touched upon issues of persistence throughout his publishing career, from "Counterparts of Persons and Their Bodies" in 1971, to "How Many Lives has Schrödinger's Cat?" in 2004. The most extensive discussions can be found in "The Paradoxes of Time Travel," "Survival and Identity," and its later postscripts, *On the Plurality of Worlds*, "Rearrangement of Particles: Reply to Lowe," and "Tensing the Copula" (2002); he remained committed to perdurance theory at every stage. In this chapter, I will explore the connections between Lewis's perdurance theory and his Humean supervenience, arguing that his influential argument about temporary intrinsics is best seen in this light. I then turn to a domestic dispute within the anti-endurantist camp: why does Lewis identify ordinary objects with world-bound parts of transworld objects, but not with time-bound parts of transtemporal objects? Given that Lewis is a counterpart theorist about modality, why isn't he a stage theorist about persistence?

15.1 Persistence and Humean Supervenience

> Humean Supervenience is yet another speculative addition to the thesis that truth supervenes on being. It says that in a world like ours, the fundamental relations are exactly the spatiotemporal relations: distance relations both spacelike and timelike, and perhaps also occupancy relations between point-sized things and spacetime points. And it says that in a world like ours, the fundamental properties are local qualities: perfectly natural intrinsic properties of points, or of point-sized occupants of points. Therefore it says that all else supervenes upon the spatiotemporal arrangement of local qualities throughout all of history, past, present and future. (Lewis 1994, 225–6)

> When philosophers claim that one or another commonplace feature of the world cannot supervene on the arrangement of qualities, I make it my business to resist. (1986b, xi)

The points of resistance are laws of nature, counterfactuals, causation, mind and language, chance, and persistence. Lewis twice discusses possible worlds exemplifying certain persistence patterns, using them as examples of distant worlds which violate Humean supervenience (he is committed to Humean supervenience only for the actual world and its neighbors):

> Two worlds might indeed differ only in unHumean ways, if one or both of them is a world where Humean supervenience fails. Perhaps there might be extra, irreducible external relations, besides the spatiotemporal ones; there might be emergent natural properties of more-than-point-sized things; *there might be things that endure identically through time or space, and trace out loci that cut across all lines of qualitative continuity*. It is not, alas, unintelligible that there might be suchlike rubbish. Some worlds have it. And when they do, it can make differences between worlds even if they match perfectly in their arrangements of qualities. (1986b, x, my emphasis)

> . . . one way to get a difference between worlds with the exact same arrangement of local qualities is to have things that are bilocated in spacetime. Take two worlds containing spheres of homogeneous matter, unlike the particulate matter of our world; in one world, the sphere spins and in the other it doesn't; but the arrangement of local qualities is just the same. These are worlds in which things persist through time not by consisting of temporal parts, but rather by bilocation in spacetime: persisting things are present in their entirety at different times. The difference between the spinning and the stationary spheres is a difference in the pattern of bilocation. (1994, 227)

The thought seems to be this: we can maintain that Humean supervenience is true at the actual world and its neighbors only if we accept that perdurance theory is true at the actual world and its neighbors. Given Lewis's determination to resist philosophical challenges to Humean supervenience, this counts in favor of perdurance theory as a description of the actual world. But how exactly would enduring objects violate Humean supervenience? Segments of the homogeneous spheres, which "trace out loci that cut across all lines of qualitative continuity," clearly violate the claim that all facts supervene upon those facts about qualitative continuity. But endurance theorists needn't be committed to the possibility of such cross-cutting objects. Here are two versions of endurance theory which seem compatible with Humean supervenience:

> **Plenitudinous endurance theory:** all and only timelike sequences of matter-filled points are trajectories of enduring objects.

> **Qualitatively-restricted endurance theory:** all and only those maximal timelike sequences of matter-filled points which exhibit appropriate qualitative and causal continuity are trajectories of enduring objects.

Each of these theories seems to be compatible with Humean supervenience. That is, each theory is such that, if the actual world and its neighbours satisfied that theory, then Humean supervenience would be true of those worlds. Any two worlds which match in their arrangements of local qualities and causal relations will also match in their distribution of (i) timelike sequences of matter-filled points and (ii) maximal timelike sequences of matter-filled points which exhibit appropriate qualitative and causal continuity. Given plenitudinous endurance theory, there is no opportunity for patterns of bilocation to become detached from lines of continuity, since patterns of bilocation are ubiquitous. Given qualitatively restricted endurance theory, enduring objects are not free to cut across lines of qualitative-causal continuity: they are somewhat like the multiply located universals which, if they exist, are compatible with Humean supervenience.

So there are versions of endurance theory which are compatible with Humean supervenience. Conversely, there are versions of perdurance theory which are incompatible with Humean

supervenience: there might be nonHumean relations between temporally separated stages. Nevertheless, the suspicion lingers that endurance is somehow inherently in tension with Humean supervenience, whilst perdurance is not. What could be the source of that tension?

Perhaps endurance is incompatible with the stronger thesis of Humean determination, according to which all facts not only supervene upon but are determined or grounded by local qualities and spatiotemporal relations. Endurance theory has it that objects wholly present at different times can stand in the relation of identity: perhaps facts about identity cannot be determined by facts about continuity, even when they supervene upon such facts.

This claim is certainly not obviously true: on the contrary, those who advocate a principle of the identity of indiscernibles are often motivated by the view that identity facts must be determined by qualitative facts. Moreover, where qualitative continuity and identity are in step with one another (as with qualitatively restricted endurance theory), it looks as if the continuity facts ground the identity facts, rather than vice versa. After all, the continuity may be explained causally, and this in turn explains the identity facts. Plenitudinous endurance theory also sits comfortably with Humean determination: a line is the trajectory of an enduring object because it is a sequence of matter-filled points.

In any case, it's unlikely that Lewis would endorse Humean determination. In 1983, he explicitly denied that brief stages are more fundamental than persisting objects:

> When I say that persons are maximal R-interrelated aggregates of person-stages, I do not claim to be reducing 'constructs' to 'more basic entities'. (Since I do not intend a reduction to the basic, I am free to say without circularity that person-stages are R-interrelated aggregates of shorter person-stages.) Similarly, I think it is an informative necessary truth that trains are maximal aggregates of cars interrelated by the ancestral of the relation of being coupled together (count the locomotive as a special kind of car). But I do not think of this as a reduction to the basic. Whatever 'more basic' is supposed to mean, I don't think it means 'smaller'. (1983a, 75)

How else might we articulate the tension between endurance theory and Humean supervenience? Isn't endurance theory somehow inherently non-local? After all, if endurance theory is true, there are relational arcs of identity connecting objects which exist at different times. But then also if perdurance theory is true, there are relational arcs of identity connecting objects which exist at different times – perduring, four-dimensional objects. No difference there.

Let's try again. If endurance theory is true, there are relational arcs of identity connecting objects which are *wholly present at different times*. But then if perdurance theory is true, there are relational arcs of non-identity connecting objects which are wholly present at different times – momentary stages of perduring objects. To the extent that cross-temporal facts about identity are temporally non-local, so too are cross-temporal facts about non-identity, that is, distinctness.

We have not yet located the conflict between endurance theory and Humean supervenience. There are versions of endurance theory which, if they were true of the actual world, would be compatible with the contingent truth of Humean supervenience, whilst there are versions of perdurance theory which are incompatible with Humean supervenience. If anything violates the spirit of Humean supervenience more generally, it may be the very distinction between perdurance and endurance theories itself, relying as it does on claims about identity and distinctness between objects wholly present at different times. Similar considerations apply to other debates involving extension and multi-location: perhaps the true Humean should resist both endurance and perdurance, both universals and tropes, and both extended and non-extended spatial simples.

15.2 In Defense of Stages

Humean supervenience in isolation does not entail perdurance theory, even for the actual world. Nevertheless, it features in Lewis's arguments for the theory. In his "Postscript B" to "Survival and Identity" (1983a), Lewis provides an argument for the existence of person-stages, an argument which can easily be generalized to non-personal stages. He begins by explaining what a person-stage is supposed to be, for the benefit of those who profess not to understand the notion: "A person-stage is a physical object, just as a person is . . . it talks and walks and thinks, it has beliefs and desires, it has a size and shape and location . . . That is what I mean by a person-stage" (1983a, 74).

Lewis then argues that such stages exist:

> First: it is possible that a person-stage might exist. Suppose it to appear out of thin air, then vanish again . . .
> Second: it is possible that two person-stages might exist in succession, one right after the other but without overlap. Further, the qualities and location of the second at its appearance might exactly match those of the first at its disappearance. Here I rely on a *patchwork principle* for possibility: if it is possible that X happen intrinsically in a spatiotemporal region, and if it is likewise possible that Y happen in a region, then also it is possible that both X and Y happen in two distinct but adjacent regions. There are no necessary incompatibilities between distinct existences. Anything can follow anything.
> Third: extending the previous point, it is possible that there might be a world of stages that is exactly like our own world in its point-by-point distribution of intrinsic local qualities over space and time.
> Fourth: further, such a world of stages might also be exactly like our own in its causal relations between local matters of particular fact.
> Fifth: then such a world of stages would be exactly like our own simpliciter. There are no features of our world except those that supervene on the distribution of local qualities and their causal relations.
> Sixth: then our own world is a world of stages . . .
> Seventh: but persons exist too, and persons (in most cases) are not person-stages. They last too long. Yet persons, and person-stages, like tables and table-legs, do not occupy spatiotemporal regions twice over. That can only be because they are not distinct. They are part-identical: in other words, the person-stages are parts of the persons. (1983a, 76–7)

Humean supervenience crops up twice in this argument. First, the "patchwork principle" of step two, whilst not strictly a consequence of Humean supervenience, is closely related to it. Then, at step five, we infer from the existence of a perdurance world which is Humeanly indiscernible from our own world to the conclusion that our own world is a perdurance world.

Endurance theorists might concede that Lewis has demonstrated the existence of sequences of person-stages, but argue that he has not demonstrated that the members of any given sequence are distinct from one another. Suppose that a spatiotemporal region is occupied by an object. Intrinsic facts about that region do not determine whether the object *also* occupies other spatiotemporal regions, either by being wholly present at more than one region, or simply by extending beyond the boundaries of the original region.

This is relevant at step two of the argument. If it is possible that a spatiotemporal region be occupied by a child, and if it is likewise possible that a spatiotemporal region be occupied by an adult, then the patchwork principle establishes that it is possible that two temporally adjacent such regions be occupied by a child and an adult respectively. But it doesn't establish that this can occur without the child's being identical to the adult, in the sense endurance theorists intend. To secure this stronger conclusion, we need the further assumption that facts about whether two adjacent regions contain the same wholly present object are not fixed by qualitative facts about those regions. And this is the illegitimate assumption that all versions of endurance theory conflict with Humean supervenience.

Even if the argument as a whole does not succeed, Lewis's final point about parts and wholes does tell against plenitudinous endurance theory, the view that there are as many enduring objects as there are time-like sequences of matter-filled points. This amounts to many, many enduring objects, one for each stage the perdurance theorist recognizes, and one for each sum of such stages. It is a short simplifying step to accept that many of these objects are partially identical to one another – related as parts and wholes – rather than entirely distinct. But to accept this is to accept perdurance theory.

Plenitudinous endurance theory is, of course, ontologically vast. But so is Lewis's version of perdurance theory: he combines a liberal approach to positing stages with a mereological universalism which generates as many persisting objects as does plenitudinous endurance theory. For Lewis, however, the persisting objects are not to be accounted as extra ontology in any robust sense: it is double-counting to count the wholes as separate from all the parts. (This accounting method will be invoked again towards the end of this chapter.)

15.3 Temporary Intrinsics

> The principal and decisive objection against endurance, as an account of the persistence of ordinary things such as people or puddles, is the problem of temporary intrinsics. Persisting things change their intrinsic properties. For instance, shape: when I sit, I have a bent shape; when I stand, I have a straightened shape. Both shapes are temporary intrinsic properties; I have them only some of the time. How is such change possible? (1986a, 203–4)

Evidently, it is possible for Lewis to have different shapes by having each shape at a different time. The problem lies in understanding what it is about the passage of time which makes it possible for Lewis to be bent at one time and straight at another, given that he could not have both shapes simultaneously.

Perdurance theorists have an immediate solution to the problem: Lewis has a temporal part which is bent, and a distinct temporal part which is straight. It is the various temporal parts which exemplify these three-dimensional shapes, whilst Lewis himself – the perduring person – satisfies the corresponding temporally relative predicates in virtue of the properties of his temporal parts. Lewis satisfies "is bent at t_1" because he has a temporal part existing at t_1 which instantiates the property *bent*.

Endurance theorists are initially offered two alternative "solutions" to the problem.

> First solution: contrary to what we might think, shapes are not genuine intrinsic properties. They are disguised relations, which an enduring thing may bear to times. One and the same enduring thing may bear the bent-shape relation to some times, and the straight-shape relation to others. In itself, considered apart from its relations to other things, it has no shape at all. This is simply incredible, if we are speaking of the persistence of ordinary things. (It might do for the endurance of [intrinsically unchanging] entelechies or universals.) If we know what shape is, we know that it is a property, not a relation. (1983a, 204)

Lewis is tantalizingly brief here, and doesn't add much when he discusses temporary intrinsics in later publications, though he asks "where have the monadic properties *bent* and *straight* gone? What have they to do with our new-found *bent-at* and *straight-at* relations…?" (2002, 4)

> Second solution: the only intrinsic properties of a thing are those it has at the present moment. Other times are like false stories…This is a solution that rejects endurance; because it rejects persistence

> altogether. And it is even less credible than the first solution. In saying that there are no other times, as opposed to false representations thereof, it goes against what we all believe. No man, unless it be at the moment of his execution, believes that he has no future; still less does anyone believe that he has no past. (1986a, 204)

Thus Lewis rejects presentism, the view that only what is present exists. Again, he is extremely brief, both here and elsewhere.

A third type of endurance-friendly "solution" is missing from *On the Plurality of Worlds*, appears in a footnote crediting Sally Haslanger and Mark Johnston in "Rearrangement of Particles," and finally gets Lewis's full attention in "Tensing the Copula." The idea is that instantiation – the having of properties – is relative to times: *having* is a three-place relation between objects, properties, and times, so *bent* and *straight* retain their status as intrinsic properties. Lewis's key objection here is that if we reify *having*, we invite Bradley's regress: in virtue of what do Lewis, *bent* and t_1 mutually instantiate *having*? There are various versions of the third strategy, and Lewis's various responses tell us plenty about his views on properties, instantiation, set-membership, and truthmaking. But endurance theorists would not need a third strategy were it not for Lewis's implacable insistence that three-dimensional shape is an intrinsic property of ordinary objects. This insistence is the core of his argument against endurance theory.

What should we make of the argument from temporary intrinsics? To be frank, it seems uncharacteristically weak. As Lewis himself shows elsewhere, the intrinsic/extrinsic distinction is not straightforward, and intrinsic/relational even less so. Moreover, relations between concrete objects and times (whatever times are), are likely to differ from paradigm relations such as *taller than*: Moorean common sense does not specify whether three-dimensional shape is such a non-paradigmatic relation. And shape is a peculiar exemplar: an object's shape may well be reducible to occupancy relations between the object and a region, together with the shape of the region itself. Finally, we shouldn't prejudge the capacity of science to surprise us about whether three-dimensional shapes are relations between their possessors, on the one hand, and other concrete objects, frames of reference or fields on the other; the same goes for mass, colour, biological species, and so on. Yet Lewis calls the argument from temporary intrinsics the "principal and decisive objection against endurance." We confront an exegetical puzzle.

Though Lewis himself does not make the connection explicit, I think that his remarks about temporary intrinsics can fruitfully be understood in the light of his Humean supervenience. If ordinary temporary properties of objects were relations between objects and times, then those relations would not supervene upon the intrinsic properties of their relata: neither objects nor times would have enough genuinely intrinsic properties to generate the rich qualitative array of the world we see around us. Would these non-supervenenient relations be amongst those exceptionally permitted by Humean supervenience? According to Lewis, these are "exactly the spatiotemporal relations: distance relations both spacelike and timelike, and perhaps also occupancy relations between point-sized things and spacetime points" (1994, 225).

The non-supervenient relations in question would not be mere occupancy relations – very many objects all occupy, or exist at, any given time, yet on this view these objects all stand in different "qualitative" relations to that particular time (conversely, a single object occupies many different times, yet stands in different "qualitative" relations to those different times). Nor are the non-supervenient relations comparable to distance relations, which relate points to points, or objects to objects, not objects to times; nor are they like relations of orientation, betweenness, or other spatiotemporal relations. So "temporary intrinsics as relations" would violate Humean supervenience.

Moreover, this explains why Lewis doesn't consider empirical objections to the view that three-dimensional shape is intrinsic: he does not aspire to defend Humean supervenience against scientific

challenges. "Really, what I uphold is not so much the truth of Humean supervenience as the *tenability* of it. If physics itself were to teach me that it is false, I wouldn't grieve" (1986b, xi; cf. 1994, 226). Lewis acknowledges quantum entanglement as a scientific challenge to Humean supervenience, and likewise he can accept empirical claims that, for example, shape or mass might be a non-supervenient relation between objects and frames of reference, or between objects and fields. His commitment is to resisting "philosophical" suggestions of non-supervenient relationality: attempts to impose an endurance-friendly account of temporary properties are philosophical in this sense, so Lewis must resist them.

Recognizing "temporary intrinsics as relations" as a violation of Humean supervenience also explains Lewis's insouciance about a relational feature of his own view. Given perdurance theory, the temporary properties of persisting objects are relations between those objects and their temporal parts. Lewis is bent at t_1, but Lewis himself is four-dimensional, and does not instantiate three-dimensional shape properties like *bent*. He is bent at t_1 only in virtue of the relation between Lewis and his temporal part Lewis-at-t_1, which does instantiate the intrinsic property *bent*. "Simply incredible"? Apparently not. The acceptability of this perdurantist relational view of change may be explained by its compatibility with Humean supervenience: an object's relations to its own parts are amongst its intrinsic features, and *bent* is an intrinsic property of the temporal part in question.

So Lewis's objection to "temporary intrinsics as relations" may be grounded in his Humean supervenience. And temporary intrinsics may be the ultimate source of the conflict between endurance theory and Humean supervenience, a conflict which we earlier struggled to locate. As we saw, neither qualitatively restricted nor plenitudinous endurance theory is committed to non-supervenient relations between different points. Instead, non-supervenience is a feature of the relations between an enduring object and the succession of points it occupies, given that the object changes in its properties over time, and given the temporary-intrinsics-as-relations account of such change. If this is a problem, it is a problem for all versions of endurance theory which permit widespread change, including plenitudinous and qualitatively restricted endurance theory; the only possible solutions, in Lewis's eyes, are to reject Humean supervenience, or else to explore other options such as presentism, or the relativization of instantiation.

There is a lot to be gained from viewing Lewis's discussion of temporary intrinsics against the backdrop of his Humean supervenience. But some unresolved issues remain. First, there is the simple point that Lewis himself does not express his concern about temporary intrinsics as arising from his Humean supervenience. Second, there is the vehemence of Lewis's rejection of temporary intrinsics as relations – "simply incredible . . . if we know what shape is . . . " – given his relatively modest goal of upholding the mere tenability of Humean supervenience. This goal requires only a demonstration of the tenability of a Humean-friendly account of change (perdurance theory, for example), not a decisive crushing of non-Humean accounts of change such as endurance theory. So I will stop short of claiming that this is what Lewis had in mind. Nevertheless, making this connection enables us to strengthen the uncharacteristically weak argument from temporary intrinsics, and to pin down the elusive tension between endurance theory and Humean supervenience.

15.4 Stages, or Sums of Stages?

Lewis's treatment of temporary intrinsics in *On the Plurality of Worlds* forms part of a discussion of both persistence and its modal analogue, for which we have no neutral term. For Lewis, possible worlds are concrete realities; amongst the parts of nearby non-actual worlds are concrete objects which closely resemble actual objects. Consider possible-Katherine, a non-actual thing which closely

– but not exactly – resembles actual-Katherine. Lewis develops an argument from accidental intrinsics which parallels his argument from temporary intrinsics, concluding that no concrete object is wholly present in more than one world: actual-Katherine and possible-Katherine are not literally one and the same thing, because they do not share all their intrinsic properties. Thus the modal analogue of endurance theory is rejected.

So what am I, Katherine, this person? There are two options. Either I am actual-Katherine, or I am the sum of actual-Katherine and all the possible-Katherines, a transworld individual with a part in each of many worlds. Lewis accepts that both these objects exist: his mereological universalism (belief in arbitrary sums) guarantees the existence of the transworld object, given the existence of the various worldbound objects. But Lewis identifies ordinary objects with worldbound individuals: Katherine is actual-Katherine. Actual objects have non-actual counterparts, and the ordinary features of those counterparts ground the truth or falsity of claims about the modal features of ordinary objects. I could have been a contender, because I have a counterpart, a possible Katherine, who is a contender in the world she is a part of.

Lewis often exploits analogies between time, space and modality (e.g. 1983b, xi), and he rejects both endurance theory and its modal analogue. But the analogy is not complete: for Lewis, ordinary objects are worldbound but they are not timebound. Perdurance theory is challenged by the stage theory of persistence, which identifies ordinary objects with brief stages instead of transtemporal sums of those stages; claims about the past and future of a given timebound stage are made true by counterpart stages existing at different times, just as modal claims about worldbound objects are made true by counterparts existing at different worlds. Stage theory is the temporal analogue of Lewis's modal counterpart theory. So why isn't Lewis a stage theorist?

Lewis doesn't tell us directly why he rejects stage theory. Instead, he assumes that ordinary objects are transtemporal sums of stages (once endurance theory has been defeated), then provides several reasons why, even given this assumption, we should accept that ordinary objects are worldbound. Identifying ordinary objects with transtemporal and transworld objects seems to be the default option, to be accepted in the absence of countervailing reasons, and Lewis provides such countervailing reasons only in the modal case. Nevertheless, examining these reasons provides some insight into the relative merits of perdurance theory and stage theory.

Lewis's "simplest" point is that most of us do not believe in transworld individuals, since we do not believe in non-actual concrete worlds. So identifying ordinary objects with actual-worldbound individuals provides at least one small respect in which Lewis can conform to "common opinion" (1986a, 220). He doesn't explore the temporal parallel, but there are both philosophers and normal people who deny the existence of the future: if this fact carries any argumentative weight at all, it favors stage theory.

A different argument seems to apply to people, but not to other ordinary objects. Lewis points out that person-stages have a special attitude towards "their" other-time stages, but not towards their other-worldly counterparts. "To some extent, stages want to fulfil the remembered desires of earlier stages . . . To a greater extent, stages want to fulfil the foreseen desires of later stages: that is prudence." In contrast "[m]y this-worldly self has *no* tendency to make the purposes of its other-worldly counterparts its own" (1986a, 219). This is because there is no transworld causation, and because there is no contingency in the collective fortunes of one's counterparts.

It is clear that transtemporal relations amongst personal stages differ from transworld relations in the way Lewis indicates. But the discussion brings out a difficulty for perdurance theory, not an advantage. Stages instantiate ordinary temporary properties, psychological and otherwise: this is key to Lewis's account of temporary intrinsics, and it is also important to his treatment of self-locating beliefs in "Attitudes De Dicto and De Se." When you lie awake at night wondering what time it is, your stage wonders where *it* is located in time, which of many wakeful stages it is.

So, as Lewis admits:

> . . . my view makes it fair to think of the desires as belonging in the first instance to my present stage, and derivatively to the persisting sum of many stages. And I agree that what my present stage wants is not to finish the book itself – it's a sensible stage so it knows that can't be expected. But that's not to say that it only cares what happens, never mind how. There is a middle ground. My present stage wants to the book to be finished in the fulfilment of its present intentions – there's the egocentric part – and that will happen only if the proper sort of causal continuity binds together my present stage with the one that finishes the book. (1986a, 126, fn. 8)

Unlike stage theorists, perdurance theorists must distinguish the instantiators of ordinary temporal properties – *desiring*, *wondering*, *being bent*, *being straight*, *being alive* – from the instantiators of sortal properties like *being a person*. This counts against perdurance theory.

Still, Lewis is right that stages' concern for "their" other-timely stages is mediated by distinctive causal connections which are absent in the transworld case. The point generalizes: we care more about actual goings-on, regardless of whether they affect us personally, than we do about other-worldly matters, even though all are supposed to be equally real, equally concrete (1986a, section 2.6). Yet the thin, acausal nature of transworld relations does not directly show that people or other ordinary objects are not transworld entities, just as the richness of transtemporal relations does not directly show that ordinary objects are transtemporal; more work is needed to bring out the connection.

The detail of Lewis's argument that ordinary objects are worldbound is set out in *On the Plurality of Worlds* (1986a, 218–20), but we get to the heart of it as early as 1971:

> We may draw an analogy between the [transtemporal] relations among stages of persons . . . and the personal . . . [modal] counterpart relations among [persisting] persons . . . But the analogy is imperfect. The [transtemporal] relations are equivalence relations, at least for the most part and as a matter of contingent fact . . . [but modal] counterpart relations are not equivalence relations. (1971, 52)

Both temporal and modal counterpart relations are reflexive, and, let us assume, symmetric. But transtemporal counterpart relations are typically transitive, whilst modal counterpart relations are not. That is, there are cases in which a is a counterpart of b, b is a counterpart of c, and yet a is not a counterpart of c. "Like all similarity relations on a sufficiently variegated domain, [modal counterpart relations] fail to be transitive because chains of little differences add up to big differences" (1971, 52).

Lewis's argument has two main steps. First, he must show that modal counterpart relations are just similarity relations, unlike temporal counterpart relations. This establishes that modal counterpart relations are not transitive. Second, he must show that the non-transitivity of modal counterpart relations establishes that ordinary objects are worldbound, whilst the few exceptions to transitivity permitted by the temporal counterpart relations do *not* establish that ordinary objects are timebound stages. I will discuss these two main steps in turn.

In *On the Plurality of Worlds* (1986a, 218), Lewis shows us that temporal counterpart relations are constrained in ways which do not apply to modal counterpart relations. First: the temporal stages of ordinary objects are bound together by causal relations. Mereological universalism commits Lewis to very many transtemporal sums, most of which are not causally coherent. But those sums of stages which correspond to ordinary objects – people and puddles – do exhibit causal continuity. In contrast, modal counterpart relations cannot involve causal continuity, for there is no transworld causation.

Second: times are ordered by the *earlier-than* relation (setting aside the complications of special relativity), and this in turn orders the stages. Thus, when we consider whether some stages are stages of the same ordinary object, we can put them in temporal order, then ask whether their collective qualitative variation over time is incremental. In contrast, the possible worlds are not set out in a neat row, and so a collection of worldbound individuals cannot be given a unique ordering. This means that modal counterpart relations cannot involve facts about "the objects in-between" two putative counterparts.

These two points together do not quite demonstrate that modal counterpart relations are just similarity relations, but they make this very plausible by ruling out two natural alternatives: after all, what else could modal counterpart relations be? So whether or not one object is a counterpart of another depends entirely on their degree of similarity in various respects, respects determined by the relevant counterpart relation. Since logical space is as variegated as can be, such similarity relations are not transitive.

This brings us to the second main step of Lewis's argument. If we accept that modal counterpart relations are similarity relations, and thus intransitive, how does this show that ordinary objects are worldbound? Intransitivity clearly shows that counterpart relations are not identity relations, because identity is transitive. But we have already accepted that possible worlds do not share parts, that counterparts are not literally identical with one another. Moreover, intransitivity does not rule out the existence of transworld individuals; these are mandated by Lewis's mereological universalism. Instead, intransitivity tells against the identification of transworld individuals with ordinary objects, by undermining the ordinary ways in which we count ordinary objects.

Suppose for the sake of argument that Katherine is a transworld individual. Exactly which transworld individual? Katherine is a maximal sum of worldbound individuals, including actual-Katherine, mutually linked by the "same person as" counterpart relation. But intransitivity ensures the existence of an enormous number of these maximal sums.

How so? Consider two personal counterparts of actual-Katherine, possible-Katherine-α and possible-Katherine-ω. Each differs from actual-Katherine in significant but different ways, and these significant differences add up, so the two are not personal counterparts of one another (they exemplify the intransitivity of counterparthood). Then actual-Katherine and possible-Katherine-α are both parts of one transworld individual, whilst actual-Katherine and possible-Katherine-ω are both parts of a second transworld individual. Each of these is an equally good candidate to be the transworld Katherine, with nothing to choose between them. And there are many more such candidate Katherines.

If people are transworld individuals, then we have a population crisis in the actual world. I seem to be alone in my office as I write this, but in fact a multitude of transworld people are here right now, typing these words. Moreover, proper names for people – and other ordinary objects – are massively ambiguous. To recover the ordinary verdict that there is just one person in my office right now, we must insist that we can count people by counting their worldbound components, or that we needn't count all the people, or that we count all the people but do not count them by identity. This choice of evils constitutes the case against identifying ordinary objects with transworld individuals.

So why the difference for transtemporal individuals? Temporal counterpart relations are richer than mere similarity, but the additional requirements of causal continuity and incremental change do not guarantee their transitivity. There are possible situations – familiar from science fiction – in which a present stage bears the temporal counterpart relation to each of two different stages existing at the same future time, stages which are not counterparts of one another: "fission" is possible. Ditto for past times: "fusion" is possible. If fission lies in my future, then there are two perduring people in my office right now, and we must explain away the common sense verdict that there is just one person here.

But such violations of transitivity are rare or nonexistent in the actual world:

> We get by because ordinary [temporal] cases are not pathological. But modality is different: pathology is everywhere . . . So the modal case will always, or almost always, give us the same choice of evils about how to count [ordinary objects] that the temporal case gives us only in connection with far-fetched stories. (1986a, 219)

For contingent reasons, perdurance theory rarely if ever forces us to reconstrue the ordinary ways in which we count actual ordinary objects (people, at least). But stage theory goes one step better: even if "fission" will occur, we do not need to choose between counting stages and counting ordinary objects, for stages just are ordinary objects. So concerns about intransitivity and counting favour stage theory over perdurance theory, just as in the modal case. But because temporal pathology is rare, this advantage is small, and may easily be outweighed if stage theory has other disadvantages. As I noted above, Lewis seems to take perdurance theory as the default view once endurance theory has been excluded, and this small advantage for stage theory may not be enough to undermine perdurance theory's status as default.

The story doesn't end there, however, for Lewis's mereological universalism and attitude to indeterminacy means that he too must reconstrue the ways in which we count ordinary objects, even in nonpathological situations. Most ordinary objects have vague spatial and temporal boundaries: there are stages around the end of my life such that it is indeterminate whether those stages are temporal parts of me. For Lewis, this amounts to indeterminacy in which of many slightly different precise objects satisfies "is a person," which one bears my name.

> Do you think you are one person? – No, there are many aggregates of temporal parts, differing just a little at the ends, with equal claim to count as persons, and equal claim to count as you. Are all those equally good claims good enough? If so, you are many. If not, you are none. Either way we get the wrong answer. For undeniably you are one. (1993, 165)

Ultimately, Lewis is untroubled by this multiplicity because the person-candidates overlap very significantly, sharing most of their parts: they are "partially identical," indeed "almost identical." (I mentioned a similar move in section 15.2: the ontological extravagance of perdurance theory is reckoned less serious than the ontological extravagance of plenitudinous endurance theory.) This reconciliation isn't available where multiplicity arises from fission or fusion cases, since the objects in such cases differ significantly from one another; all the more so in the universal pathology of the modal.

Not all of us share Lewis's preference for perdurance theory over stage theory, and of course not all of us follow him in rejecting endurance theory. But his treatment of this issue – which barely reaches the surface of his writing – is an especially beautiful example of Lewisian metaphysical and semantic views working to support one another. It illustrates the systematicity of his metaphysics, integrating the rejection of worldly indeterminacy and the quasi-identification of parts and wholes with headline issues about worlds, times, and properties. And it offers a masterclass in the art of recognizing both the power and the limitations of analogy.

Acknowledgments

Many thanks to Jonathan Schaffer for his comments on an earlier draft. This work was supported by the European Community's Seventh Framework Programme FP7/2007-2013, under grant agreement no. FP7-238128.

References

Lewis, D. (1971). Counterparts of persons and their bodies. *Journal of Philosophy*, 68, 203–11.
Lewis D. (1983a). Postscripts to "Survival and identity." In Lewis (1983b), pp. 73–7.
Lewis, D. (1983b). *Philosophical papers*, Vol. I. Oxford: Oxford University Press.
Lewis, D. (1986a). *On the plurality of worlds*. Oxford: Blackwell.
Lewis, D. (1986b). *Philosophical papers*, Vol. II. Oxford: Oxford University Press.
Lewis, D. (1993). Many, but almost one. In K. Campbell, J. Bacon & L. Reinhardt (Eds.), *Ontology, causality and mind: essays on the philosophy of D.M. Armstrong*, pp. 23–37. Cambridge: Cambridge University Press. Reprinted in Lewis (1999), pp. 164–82. Page references are to the latter version.
Lewis, D. (1994). Humean supervenience debugged. *Mind*, 103, 473–90. Reprinted in Lewis (1999), pp. 224–47. Page references are to the latter version.
Lewis, D. (1999). *Papers in metaphysics and epistemology*. Cambridge: Cambridge University Press.
Lewis, D. (2002). Tensing the copula. *Mind*, 111, 1–13.

16

"Perfectly Understood, Unproblematic, and Certain"

Lewis on Mereology

KAREN BENNETT

16.1 Four Theses about Composition

David Lewis famously takes mereology "to be perfectly understood, unproblematic, and certain" (1991, 75). It is central to his thought, appearing in his discussions of set theory, modality, vagueness, structural universals, and elsewhere. He held views not only about how composition works and when it occurs, but also about the role of mereology in philosophy. In this chapter, I will proceed by articulating four theses that Lewis holds about composition. (I would call them the four Us, if only 'unguilty' were a word!) Three of them are familiar; Lewis himself explicitly articulates and relies upon them. The fourth remains implicit, but it is nonetheless important. Here they are:

Composition is *unique* – the same things cannot have two different fusions.
Composition is *unrestricted* – any two things whatsoever have a fusion.
Composition is *ontologically innocent* – composed entities do not "count" beyond their parts.
Composition is *unmysterious* – it is not problematic to treat it as primitive, and can function in demystifying explanations.

I will devote a section to each thesis: explaining what it says, pointing to the texts that illustrate that Lewis believes it, and explaining why Lewis believes it. These sections are largely expository. But woven in between them are interstitial sections in which I reflect upon further questions that arise, and draw further lessons.

Two preliminary caveats. First, I will often talk of 'things' or 'objects' either having parts or being parts of something else. This is not meant to suggest that only concrete physical objects stand in the parthood relation. Lewis allows that things that do not occupy space and time can also do so (1991, 75), and I will follow him in this. My choice of words is just that – a choice of words, intended to be neutral about just what kinds of things can be part of others.

Second, the four theses around which this piece is structured do not exhaust Lewis's claims about mereology. He also believes, for example, that parthood is transitive (e.g. 1991, 74) and that atomless

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

gunk – entities all of whose parts have proper parts – is possible (e.g. 1986a, 30; 1991, 20–1). I have simply chosen the four theses because they are particularly interesting and revealing.

16.2 First Thesis: Uniqueness

The claim that composition is unique is the claim that if some entities compose something, they compose exactly *one* thing. You can never build two distinct composites out of the same parts. That is, sameness of parts is sufficient for identity: if *x* and *y* have the same parts, then 'they' are identical. And although Lewis only explicitly mentions and relies upon that conditional (1986a, 36; 1991, 74, 78, 100), the converse is guaranteed by Leibniz's Law. Indeed, in formal mereology, the idea is usually expressed as a biconditional, called the extensionality of proper parthood: any items that have proper parts at all are identical just in case they have all the same proper parts.

Uniqueness (or extensionality) is a powerful claim. It entails, for example, that the very same parts cannot compose both a lump of clay and a distinct statue. Consequently, those who believe that constitution is not identity either deny extensionality (e.g. Baker 2002) or else claim that, contra appearances, coincident entities are not composed of exactly the same parts after all (e.g. Koslicki 2008).

Lewis uses the label 'uniqueness' and explicitly endorses the thesis in *Parts of Classes* (1991, 74). It is important to his treatment of set theory in terms of mereology and the relation of singleton formation; for more detail, see Burgess, chapter 29, this volume. But another important appeal to the uniqueness of composition occurs in his dispute with David Armstrong over the nature and existence of structural universals.

Armstrong believes that properties are universals, wholly located in each of their instances, and that some universals are structural. A structural universal is a universal such that anything that instantiates it does so in virtue of the fact that its *parts* instantiate other universals, relational and otherwise. To see the idea here, consider a methane molecule: it consists of four hydrogen atoms bonded to a carbon atom in a particular way. On Armstrong's view, the universal *methane* itself is structured in a way that echoes the compositional structure of each particular methane molecule.

Lewis is deeply suspicious of structural universals (1986a). For present purposes, all that matters is his criticism of what he calls the 'pictorial conception' of structural universals, by analogy with what he calls 'pictorial ersatzism' about possible worlds (1986c, 165–74). On this line, a structural universal is literally composed of other universals; they have other universals as parts. Lewis argues:

> Each methane molecule has not one hydrogen atom but four. So if the structural universal *methane* is to be an isomorph of the molecules that are its instances, it must have the universal *hydrogen* as a part not just once, but four times over. Likewise for *bonded*, since each molecule has four bonded pairs of atoms. But what can it mean for something to have a part four times over? What are there four of? There are not four of the universal *hydrogen*, or of the universal *bonded*; there is only one. (1986a, 34)

The clear hidden premise here is uniqueness (see also 1986b, 92). Without it, the three universals *hydrogen*, *carbon*, and *bonded* could compose any number of structural universals – all of the vast array of hydrocarbons. But *with* it, they can compose only the straightforward mereological fusion of the three simpler universals. (For an alternative mereology that allows things to have parts twice over, see Bennett 2013.)

Armstrong's reaction to Lewis is interesting. Initially, he not only denies the principle that two distinct things cannot be composed of the same parts, but claims to *counterexample* it (1986). He points out that the state of affairs of John's loving Mary is not the same as the state of affairs of

Mary's loving John, despite the fact that they have the same three constituents (John, Mary, and the *loving* relation). States of affairs, he says, violate uniqueness just as structural universals do. He therefore suggests not just that Lewis is wrong, but that he has in fact begged the question against him (85).

Later, however, under Lewis' "benign prodding" (1988, 312), he came to agree that neither states of affairs nor structural universals are literally composed of parts at all. Instead, he claims their constituents make them up in a different way, by means of a *sui generis* relation of "non-mereological composition" (1988, 1997). That is, Armstrong first claims that ordinary composition violates uniqueness, and later backs down to the claim that there are *other* generative relations that do. Yet he never explains why he changed his mind; in fact, neither he nor Lewis ever acknowledges the shift at all. And they both take it for granted that the later move is better (though Lewis remains mystified, of course; 1986b, 92).

This sheds some light, I think, on the question of why Lewis believes that composition obeys uniqueness in the first place. Lewis does not *argue* against the claim that distinct things can be composed of the same parts; he simply calls it "unintelligible" (1986a, 36). Clearly, part of his implicit motivation is that he wants mereology to echo set theory. But, equally clearly, part of it is that he honestly does find the denial of uniqueness unintelligible. That is, he takes uniqueness to be true as a piece of *conceptual analysis*. Just as nothing can count as a square unless it has four right angles, nothing can count as parthood or composition if it does not obey uniqueness. This explains his apparent thought that Armstrong simply *must* be appealing to some other relation. He can call it whatever he likes, but it isn't composition. I will return to this point in section 16.4.

16.3 Second Thesis: Unrestricted Composition

The claim that composition is unrestricted is the claim that any things whatsoever have a fusion. It – or, at least, the claim that it is necessary that composition is unrestricted – entails that it is not possible for there to be only two things. If *a* exists and *b* exists, there is guaranteed to be some *c* that they compose. This entails that there is a fusion of, say, this chair, the moon, and a tree in Madagascar. Combined with Lewis's eternalism, it also entails that there are cross-time sums – go ahead and add Abraham Lincoln or a woolly mammoth to that previous fusion. Combined with Lewis' perdurantism, it also entails that there are cross-time sums of instantaneous temporal parts – there is an object composed of alternating temporal parts of you and that woolly mammoth. Combined with Lewis's modal realism, it entails that there are transworld sums – there is an object composed of any one of the previous fusions plus a flying pig. Finally, since Lewis sees no reason to restrict composition to concrete objects (1986a, 212, n. 9; 1991, 7), it also entails that there is something composed of any of the previous fusions plus the number 7. We don't *talk* about most of these weird entities, of course; we implicitly restrict our quantifiers to range over more ordinary things. But they exist nonetheless, in just the way that you and I do. We must acknowledge the strange fusions when we speak with our "quantifiers wide open" (Lewis 1991, 80; see also 1986c, 2–3).

Lewis endorses unrestricted composition in both *On the Plurality of Worlds* (1986c, 211) and *Parts of Classes* (1991, 74). He relies upon it to characterize set theory (1991, 101) and to argue that counterpart theory can be reformulated in terms of transworld individuals (1986c, 213–17). So why does he believe it? Partly because he sees no good reason to deny it (1991, 19), partly because it preserves the analogy between set theory and mereology, but mostly on the basis of what has come to be known as the "argument from vagueness" (1986c, 212–13; 1991, 80–1).

Suppose composition is restricted "in accordance with our intuitions about this-worldly cases" (1986c, 212) – suppose, for example, that there are chairs, but not chair–moon fusions. But Lewis

claims that any way of restricting composition so all and only "ordinary" composites exist will be hopelessly vague. There won't be a sharp line between cases where composition occurs and cases where it does not:

> it's not on to say that somewhere we get just enough contrast with the surroundings, just enough cohesion . . . to cross a threshold and permit composition to take place, though if the candidate class had been just a little worse it would have been sumless. (1986c, 212)

But if it is vague just how cohesive (for example) some things have to be to compose something, then it is vague when composition occurs. And if composition is vague, then so is *existence*. That is, if it is vague whether *a* and *b* compose anything, it is vague whether *there is* a thing that they compose – it is vague whether that composite exists. But that, says Lewis, cannot be; "the only intelligible account of vagueness locates it in our thought and language . . . vagueness is semantic indecision" (212). Lewis's example here is the outback. There is no determinate answer to the question, "Where exactly does the outback begin?" But, according to Lewis, that is not because

> there's this thing, the outback, with imprecise borders; rather there are many things, with different borders, and nobody has been fool enough to try to enforce one of them as the referent of the word 'outback'. (212)

In short, working backwards: there are no vague objects, which entails that existence is not vague, which entails that composition is not vague, which entails that composition is not restricted in any way. Lewis's claim is that only the large ontology generated by unrestricted composition enables a coherent treatment of vagueness.

For more careful discussion of the argument from vagueness, and an extension of the argument to justify four-dimensionalism, see Sider 2001, chapter 4, §9. For recent defense of the idea that there *can* be indeterminacy 'out in the world', rather than in our thought and language, see Barnes and Williams 2011. For a classic book-length defense of an alternative to unrestricted composition – and careful articulation of the 'special composition question' to which it is an answer – see van Inwagen 1990.

16.4 First Reflection: Motivating Mereological Principles

Notice that Lewis believes uniqueness and unrestricted composition for very different reasons. He has a substantive philosophical argument for unrestricted composition – the argument from vagueness. But, as I noted in section 16.2, he says little to nothing about why he believes uniqueness. It certainly appears that he believes it largely on the basis of conceptual analysis – reflection on the notion of composition is supposed to reveal that the same entities can only compose one whole.

This difference is worth noting, and not just as a piece of Lewis scholarship. After all, both theses are valid in classical extensional mereology. (Unrestricted composition is an axiom; the extensionality of parthood is a theorem that follows from two axioms: anti-symmetry and strong supplementation.) So there is a question, independent of Lewis, of *why* they are valid in that formal system – that is, of why principles that validate them have been chosen as axioms. I do not mean this as a history-of-philosophy question about the intentions of people like Lesniewski or Leonard and Goodman; I mean it as a question about why they are *usually* taken as axioms, about why a system that does so is widely accepted, or at least very familiar.

The answer to this question is that there is no across-the-board answer to this question. Lewis is not alone in having different motivations for the principles; his doing so simply echoes a more general fact that reveals something about the nature and point of formal mereology.

Formal mereology is a somewhat odd beast. It is in part simply a codification of our ordinary notions of parthood and composition – just conceptual analysis in fancy, symbolic garb. But that is not all it is. Conceptual analysis plausibly secures the anti-symmetry of parthood, and that it obeys weak supplementation. It is somewhat less plausible to claim that conceptual analysis secures the uniqueness of composition. But conceptual analysis clearly does *not* secure the hotly disputed claim of unrestricted composition. That is a substantive and controversial claim about what exists, not guaranteed by the meanings of terms like 'part' and 'compose'.

In short, mereological systems do not simply make analytic claims about the nature of parthood, but also reflect other philosophical commitments and purposes. Consequently, both in Lewis and more broadly, different principles can be motivated in quite different ways – and are therefore susceptible to quite different sorts of challenges.

16.5 Second Reflection: Persistence

Regardless of how they are motivated, the first two theses might appear to generate odd consequences about persistence. One such consequence apparently follows from uniqueness alone; another apparently follows from uniqueness and unrestricted composition taken in conjunction.

First, the putative consequence of uniqueness alone – it appears to entail mereological essentialism, the claim that composites cannot survive the gain or loss of parts. Here is the rough train of thought. According to uniqueness, sameness of parts is necessary for identity. Thus in order for a composite x at time t_1 to be identical to composite y at t_2, x at t_1 and y at t_2 must have exactly the same parts – that is, x must have all the same parts at t_2 as at t_1.

Second, the putative consequence of uniqueness and unrestricted composition taken together – they appear to entail that no composite object can be destroyed unless at least some of its parts are destroyed as well. Here's why. Suppose I smash a cup to smithereens with a hammer. Well, the smithereens still exist, so by unrestricted composition they still compose something. That is, there is a thing composed of exactly the same parts that once composed the cup. Since uniqueness says that sameness of parts is sufficient for identity, it follows that the cup is identical to that scattered thing. That renders my smashing rather ineffective; I didn't destroy the cup, but only spread it out over a larger spatial region. In short, uniqueness and unrestricted composition together seem to entail that if an object is composed of the *xx*s, it exists whenever the *xx*s do, no matter what the relations among the *xx*s.

Both of these arguments are too quick; both putative consequences can be blocked. The key point is that there is no direct entailment from the mereological theses to any claims about how composites persist through change. Any such entailment depends upon background views about the nature of time and persistence.

For example, an eternalist perdurantist – like Lewis himself (see Hawley, chapter 15, this volume) – has replies to both arguments. Suppose in particular a version of the view according to which persisting objects are spacetime 'worms' (as opposed to the stage view of Hawley 2001; Sider 2001) – cross-time composites of momentary time-slices, which are themselves composed of spatial parts. On this view, an ordinary persisting composite like a cup has three kinds of parts that are worth distinguishing. First, the cup has *spatial parts*. Many of these, like electrons, are persisting entities, themselves composite spacetime worms. Second, the cup has *temporal parts* – time-slices that exist only for one moment. Third, the cup has spatially smaller momentary parts, which could equally well

be characterized as temporal parts of spatial parts or as spatial parts of temporal parts. (Consider a proper part *p* of a temporal part *t* of the cup. Since *t* has no proper temporal parts of its own, *p* must be a momentary entity, spatially smaller than the cup. It is a (non-persisting) spatial part of the temporal part of the cup; it is also a temporal part of some (persisting) spatial part of the cup, like an electron.)

Given this picture, an ordinary composite object like a cup can perfectly well have different spatial parts at different times. For the cup to have an electron as a part until *t*, and then not have it thereafter, is for it to have *part* of the spacetime worm that is the electron as a part. It is for the cup's temporal parts up until *t*, and only until *t*, to have the electron's temporal parts until *t* as parts. This does not violate uniqueness. To see this, consider a temporal part *x* of the cup from the interval when the cup has the electron as a part, and a temporal part *y* of the cup from the interval when the cup does not have the electron as a part. Those short-lived entities *x* and *y* have different parts, and uniqueness entails that *they* are distinct. But that is not at all the same as saying that the cup goes out of existence when it loses the electron. The cup is a temporally extended entity that has a (temporal) part composed of some things, and another (temporal) part composed of some other things; this does not violate uniqueness any more than does the fact that my elbow has different parts than my knee.

A similar move blocks the apparent commitment to the claim that composite objects survive the massive rearrangement of their parts. Suppose that at t_2 I smash the cup into smithereens. Each smithereen itself persists through time, and therefore has temporal parts of its own. Call the t_1 time slices of the smithereens the *xx*s. Call the t_2 time slices of the smithereens the *yy*s. Unrestricted composition entails that at t_1, the *xx*s compose a temporally unextended spatially unified entity *y*; it also entails that at t_2 the *yy*s compose a temporally unextended scattered object *z*. It further entails (given the eternalism currently being assumed) that there is a cross-time composite *c* of which both *y* and *z* are parts. However, neither unrestricted composition nor uniqueness entail that *y* is identical to *z* – *y* and *z* are momentary time-slices, not persisting entities – nor that *c* is the *cup*. The cup is not identical to *c*; it has *y*, but not *z*, as a part. The cup ceases to exist when it is smashed.

Now, there are two important points to notice about these perdurantist replies. First, they deny neither that fusions have their parts essentially (on that, see Uzquiano 2014), nor that fusions exist whenever their parts do, regardless of how those parts are arranged. The perdurantist rather claims that ordinary objects are fusions of temporal parts rather than merely of spatial ones. As a consequence – this is the second point – these replies do not touch the related worries that ordinary persisting objects have their *temporal* parts essentially (see van Inwagen 1981) and are guaranteed to exist if their *temporal* parts do. If desired, these modal worries can be blocked with additional machinery; Lewis himself would invoke counterpart theory here. (See Beebee and MacBride, chapter 14, and Ismael, chapter 12, this volume, for more detail.)

What about endurantism? Endurantists think that objects persist through time by being wholly present at each moment at which they exist, and therefore take the persistence of a composite across time to involve the strict numerical identity of a composite that exists at one time with a composite that exists at another. Matters are more complicated here, and I will not go into great detail. Suffice it to note that, first, the concern about mereological essentialism is just an instance of a broader question the endurantist faces about how to reconcile Leibniz's Law with *any* change over time. (See Lewis 1986c, 202–4 for an initial statement of the problem; 1988 for further discussion; Haslanger 1989 for an endurantist solution.) Many endurantists will index property instantiation – or the possession of parts – to a time. One way to do this is to say that the cup has-at-t_1 the *xx*s as parts, and has-at-t_2 the *yy*s as parts. What about the second concern, about whether uniqueness and unrestricted composition entail that the cup exists as long as its parts do, no matter how scattered? The endurantist who accepts uniqueness and unrestricted composition does appear to be committed to

the claim that the spatially unified object y that fuses the smithereens at t_1 persists as the spatially scattered object z that fuses the smithereens at t_2. Perhaps there is room for her to deny that ordinary objects like cups are mereological fusions of spatial parts. Unlike the perdurantist, however, this requires her to deny that they are mereological fusions *at all*, and thus leaves her with the question of what they are instead.

I obviously do not intend to resolve these issues here. I make only the following two claims in this section. First, uniqueness *alone* does not entail mereological essentialism, and the combination of uniqueness and unrestricted composition *alone* does not entail that composite objects persist through all kinds of rearrangement and separation of their parts. Deriving any such claims requires supplementing the mereological principles with claims about how persistence itself works. Second, on Lewis's own views about time and persistence (and modality), the problematic consequences do not follow.

For more on Lewis's views of time and persistence, see Hawley, chapter 15, this volume. See Varzi (2010, §3.2) for general discussion of what uniqueness (extensionality) does and does not entail.

16.6 Third Thesis: Ontological Innocence

Many people dislike Lewis's second thesis, unrestricted composition, because it entails that so very many things exist. But Lewis thinks this is a completely wrong-headed reaction:

> Given a prior commitment to cats, say, a commitment to cat-fusions is not a *further* commitment. The fusion is nothing over and above the cats that compose it. It just *is* them. They just *are* it. Take them together or take them separately, the cats are the same portion of Reality either way . . . If you draw up an inventory of Reality according to your scheme of things, it would be double counting to list the cats and then also list their fusion . . . The new commitment is redundant, given the old one. (1991, 81–2)

This is the third thesis: composition is *ontologically innocent*. This thesis would, if true, assuage the worry that accepting unrestricted composition leads to a bloated ontology. But *is* it true? Why exactly should we think so?

If Lewis's claim were that the fusion is *literally identical* to the cats that compose it, he would clearly be entitled to ontological innocence. After all, a thing is "nothing over and above" *itself*, and should not be entered on the inventory of Reality twice. But, despite his evocative language in the above passage, that is not exactly his claim. Although Lewis endorses what he calls the "thesis of Composition as Identity" (1991, 82), he does not mean that composition *is* identity, that they are the very same relation. He rather intends the much weaker claim that composition is "strikingly analogous" (84) to identity. Indeed, Lewis explicitly denies the stronger claim (84, n. 12; see also 87). By doing so, he avoids certain difficult questions – for example, how a relation that holds many–one can be reflexive, or obey Leibniz's Law. But there are plenty of other questions that do arise. In particular, in what does the striking analogy between composition and identity consist? And how does that analogy support the thesis of ontological innocence?

Those questions are tricky because they are interwoven. On the one hand, Lewis seems to suggest that Composition as Identity *justifies* the thesis of ontological innocence: he says that "it is *in virtue of* [Composition as Identity] that composition is ontologically innocent" (82, emphasis mine). But, on the other hand, he also says that innocence is part of the *content* of Composition as Identity – it is one of five "aspects of the analogy" that Lewis lays out. Here are those five aspects (1991, 85–6):

1. Both composition and identity are *ontologically innocent*. If $x = y$, y is not an ontological commitment beyond x; similarly, if the xxs compose y, y is not an ontological commitment beyond x.
2. Both composition and identity are *automatic*. No special conditions beyond the existence of x must be met in order to guarantee the existence of a y identical to x; similarly, given unrestricted composition, no special conditions beyond the existence of the xxs must be met in order to guarantee the existence of a y composed of the xxs.
3. Both composition and identity are *unique*. No x can be identical to more than one thing; similarly, no xxs can compose more than one thing.
4. Both composition and identity involve *property inheritance*. Each x's properties supervene (trivially) on x's properties, intrinsic and relational. Similarly, each x's properties supervene on the properties of its parts, intrinsic and relational. [Lewis's own rough characterization of the fourth analogy (85) makes it sound as though he thinks that both identity and composition obey Leibniz's Law: as though he thinks that the properties of a thing are the very same as the properties of its parts. My more plausible version reflects the fact that he denies this a few pages later (87).]
5. Both composition and identity *block multi-location*. Each x is located exactly where it itself is; similarly, each x is located exactly where its parts are.

Yet, even setting aside aspect 1, the thesis of ontological innocence that is the topic of this section, this is a somewhat odd list. The entries do not all have the same status. Two of them – aspects 2 and 3 – are Lewis' controversial theses of unrestricted composition and uniqueness. As we have seen, he discusses these at some length throughout his work, and feels at least some obligation to defend them. In contrast, two of them – aspects 4 and 5 – are fairly uncontroversial and not really in need of argument. They are interesting phenomena that can legitimately be taken as data; it is hard to deny that composites inherit their properties and locations from those of their parts. Yet although these phenomena may not need to be defended, they do need to be explained. *Why* do composites inherit their properties and locations from their parts? Here, again, we would have an answer if composition literally *is* identity. But, again, we do not obviously have an answer given the weaker claim that composition is merely *like* identity. In fact – and this is the important point – we certainly do not have an answer in light of the fact that property inheritance and the ban on multi-location are given as part of what it *means* to say that composition is like identity. It would be circular to 'explain' those phenomena in terms of composition as identity.

Now, back to ontological innocence. Like the second and third aspects of the analogy, and unlike the fourth and fifth, this is an extremely controversial thesis that requires defense (e.g. van Inwagen 1994; Merricks 2001). Yet, like the fourth and fifth aspects, including it as part of the content of the analogy itself undermines the usefulness of the analogy. If the analogy between composition and identity partially consists of the claim that composition is ontologically innocent, the analogy cannot also justify the claim that composition is ontologically innocent.

So Lewis's famous discussion in *Parts of Classes* does not actually provide any argument at all for ontological innocence. Still, though, there are two arguments for ontological innocence that he could have given but did not – two non-circular arguments that are close enough to what he says that they may have been at least in the ballpark of what he intended.

First, Lewis could have argued that composition is like identity in certain respects *other than* ontological innocence, argued that those similarities justify extending the analogy, and concluded that composition is like identity in that it generates no new ontological commitments. This would be a classic argument by analogy. I do not know whether such an argument can be made compelling; I merely point out that it involves a highly non-trivial step that Lewis does not take in *Parts of Classes*.

Second, he could have defended the ontological innocence of composition *without appeal* to the thesis of Composition as Identity. Indeed, the natural beginnings of an argument can be found in the thesis of unrestricted composition, and Lewis's idea that it makes composition "automatic." If composition always occurs – if any two or more things have a fusion – then the existence of the *xx*s guarantees the existence of their fusion. No further conditions need to obtain; nothing more needs to happen for that fusion to exist. But if nothing more needs to happen for the fusion to exist, then the fusion is nothing over and above the *xx*s. Now, that last sentence contains the obvious gap in the argument, and I admit that I am not sure how to fill it. Again, however, doing so is Lewis's task, not mine. (Note that several people accept the converse of this argument-sketch – that is, the claim that Composition as Identity entails unrestricted composition. See Sider 2007, though see Cameron 2012 for some pushback.)

Additional reading suggestions: Lewis explicitly states that he is following Donald Baxter (1988a, 1988b). See Sider 2007 for detailed discussion of various composition as identity theses of varying strength. For a contemporary collection, see the forthcoming volume edited by Baxter and Cotnoir. For a non-Lewisian explication of the idea that some things – not just composites – are "nothing over and above" other things, see chapter 7 of Bennett (forthcoming).

16.7 Fourth Thesis: Unmysteriousness

The fourth and final thesis is one that Lewis never explicitly articulates or defends, but which he clearly believes: composition is *unmysterious*. It is unmysterious in (at least) the following two senses:

1. Composite entities are exactly as mysterious as their parts – no more.
2. No analysis of the composition relation in other terms is either forthcoming or required.

A few remarks about each are in order.

The key idea behind the first aspect of unmysteriousness is that the composition relation adds no further mystery to the things it acts upon. This falls out of ontological innocence, or perhaps the thesis of Composition as Identity more broadly. If a composite is nothing extra, nothing beyond its parts – if it just *is* the parts – then how *could* a composite be problematic if its parts are not? All composition does is wrap many into one. Note that this train of thought yields a strategy for rehabilitating prima facie dodgy entities: argue that they are composed of unmysterious parts, parts that the skeptic already accepts. A central example here would be Lewis's account of possible worlds. Don't fret about them, he seems to say; they are just fusions of more-or-less ordinary things, albeit causally and spatiotemporally isolated from us (1986c, 1–3, 69–71).

Lewis's commitment to the second aspect of unmysteriousness can be read off the fact that he neither offers an analysis of composition, nor even expresses worry that he hasn't got one. Indeed, I cannot think of anywhere that he discusses the issue at all. (Contrast van Inwagen's discussion of the "general composition question"; 1990, chapter 4.) Of course, he claims to know a great deal about how composition *behaves*: it obeys uniqueness, unrestricted composition . . . etc. That is presumably why he claims that it is "perfectly understood" (Lewis 1991, 75). Still, though, a list of rules that composition obeys is not an account of what composition *is*. (Compare the standard format of formal mereological systems: they introduce a primitive mereological predicate (parthood, proper parthood, or overlap), and then introduce axioms that govern how the predicate can be used.)

Both aspects of the fourth thesis can be illuminated by comparing Lewis's attitude towards composites and the composition relation to his attitude towards singleton sets and the relation of singleton formation. According to Lewis, singletons are "mysterious" (1991, §2.1) and, as for "the

generation of unit sets from their members . . . God knows what it is" (1986a, 37). Set theory can't do without singletons, unfortunately, and mathematics can't do without set theory. So Lewis accepts singleton formation as a primitive relation (though he also presents a structuralist alternative), but he clearly would have been happier if he could have reconstructed set theory purely in comfortable, easy, mereological terms (1991, preface, §2.8).

So what is supposed to be so bad about singletons and singleton formation? Perhaps the best summary is Lewis's imaginative description of a student trying to grasp the idea:

> he has no elements or objects . . . to be 'combined' or 'collected' or 'gathered together' into one . . . Rather, he has just one single thing, the element, and he has another single thing, the singleton, and nothing he was told gives him the slightest guidance about what that one thing has to do with the other. (1991, 30)

The complaint begins, then, with the fact that singleton formation is one–one, not many–one. Whatever it does, it does not simply wrap many into one (1991, 30, 41). Singletons sets are "atoms" (31), not entities with the original entities as parts. Indeed, singletons are "wholly distinct from the familiar individuals" that are their members (31, 41). This means two things. First, we know very little about what these singletons are like – they are wholly new entities magically zapped into existence. Second, the relation of singleton formation is not ontologically innocent (also 87). Indeed, we know almost as little about the relation as we do about the entities it generates: all we know is that it is not ontologically innocent, and it is not intrinsic (34). Now contrast composites and composition. No mysteries there, according to Lewis; all composition does is take us from plural to singular. The crucial difference, note, is the putative ontological innocence of composition. Composites are *not* wholly distinct from their parts, so we do know as much about their nature as we know about the nature of the parts.

Does this difference really hold up? One way to challenge it would be to argue that singleton formation is not mysterious after all. The other way to challenge it would be to argue that composition is not as . . . wholesome as Lewis thinks. This would perhaps be a good time to remember that Lewis offers no real argument for the ontological innocence of composition. Further, it is at best questionable to demystify composition by means of the claim that composites are "nothing over and above" their parts when that phrase cannot be interpreted as meaning "numerically identical to."

16.8 Fourth Reflection: Privileging Mereology

The discussion thus far – both mine and Lewis's – makes it sound as though Lewis acknowledges exactly two generative relations. He embraces composition, will tolerate singleton formation, and wants no truck with anything else, like Armstrong's non-mereological composition. Yet this is not accurate. Non-causal uses of 'because' and 'in virtue of' are liberally sprinkled throughout his work, and he heavily relies on supervenience. (This prompts an interesting question: what would Lewis have thought of the currently popular view, largely due to Kit Fine (1994) (though see also McLaughlin and Bennett 2011, §3.5), that 'in virtue of' cannot be explicated in terms of modal notions like supervenience? Here is one passage that might provide a clue about his possible reaction: " 'first' we have individuals, and the classes come 'later'. In some sense"; Lewis 1991, 9.) Even restricting my attention to supervenience yields plenty of familiar examples:

> Among worlds with no alien natural properties, mental properties supervene on physical ones. (Lewis 1983, 208–12)
> Internal relations supervene on the intrinsic natures of their relata. (1986c, 62)

Representation *de re* supervenes on qualitative character (1986c, 221).
Everything supervenes on the worldwide distribution of intrinsic properties over space–time points. (1986d, ix; 1994, 473)

What should we make of such claims?

They clearly involve the idea that one kind of property, relation, or entity gives rise to or generates another, less fundamental kind. (Again, supervenience alone does not entail any such thing (see McLaughlin and Bennett 2011, §3.5), but it is indisputably what Lewis intends.) Yet they equally clearly do not invoke composition; Lewis is not saying that mental properties have physical properties *as parts*, nor that internal relations are *made of* the intrinsic properties of their relata. So they are in some sense generative, but not mereological. Rather like singleton formation. Why is supervenience kosher when singleton formation is not?

In fact, supervenience seems to share some of the features that worry Lewis about singleton formation. It is not always many–one; consider the case of materialism. Even in cases where that description can at least sort of be pressed into service – internal relations arguably 'wrap' the intrinsic features of multiple entities into a single relation – it is not the case that the supervening features 'gather together' the subvening features in anything like a compositional sense. What about ontological innocence? The 'nothing over and above' terminology is tempting here, and widely used by others if not by Lewis (see McLaughlin and Bennett 2011, §3.4). However, it again does not mean numerical identity, so the issue remains open for further discussion.

The questions here are these: is there a broader class of generative relations – what I elsewhere call 'building relations' (Bennett forthcoming) – that includes composition among others? If so, what makes a relation count as a member of that class? Why exactly does Lewis take some generative relations to be more respectable than others? Is he right to do so? Lewis did so much to deepen our understanding of composition and its usefulness; I wish he had turned his formidable mind to these broader questions as well.

References

Armstrong, D. (1986). In defence of structural universals. *Australasian Journal of Philosophy*, 64, 85–8.
Armstrong, D. (1988). Are quantities relations? A reply to Bigelow and Pargetter. *Philosophical Studies*, 54, 305–16.
Armstrong, D. (1997). *A world of states of affairs*. Cambridge: Cambridge University Press.
Baker, L. (2002). Replies to critics. *Philosophy and Phenomenological Research*, 64, 623–35.
Barnes, E. & Williams, J.R.G. (2011). A theory of metaphysical indeterminacy. In K. Bennett & D. Zimmermann (Eds.), *Oxford Studies in Metaphysics*, Vol. 6, pp. 103–48. Oxford: Oxford University Press.
Baxter, D. (1988a). Identity in the loose and popular sense. *Mind*, 97, 575–82.
Baxter, D. (1988b). Many–one identity. *Philosophical Papers*, 17, 193–216.
Baxter, D. & Cotnoir, A. (Eds.) (forthcoming). *Composition as identity*. Oxford: Oxford University Press.
Bennett, K. (2013). Having a part twice over. *Australasian Journal of Philosophy*, 91, 83–103.
Bennett, K. (forthcoming). *Making things up*. Oxford: Oxford University Press.
Cameron, R. (2012). Composition as identity does not settle the special composition question. *Philosophy and Phenomenological Research*, 84(3), 531–54.
Fine, K. (1994). Essence and modality. *Philosophical Perspectives*, 8, 1–16.
Haslanger, S. (1989). Endurance and temporary intrinsics. *Analysis*, 49, 119–125.
Hawley, K. (2001). *How things persist*. Oxford: Oxford University Press.
Koslicki, K. (2008). *The structure of objects*. Oxford: Oxford University Press.
Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, 61, 343–77.
Lewis, D. (1986a). Against structural universals. *Australasian Journal of Philosophy*, 64, 25–46.

Lewis, D. (1986b). A comment on Armstrong and Forrest. *Australasian Journal of Philosophy*, 64, 92–3.
Lewis, D. (1986c). *On the plurality of worlds*. Oxford: Blackwell.
Lewis, D. (1986d). *Philosophical papers*, Vol. II. Oxford: Oxford University Press.
Lewis, D. (1988). Rearrangement of particles: reply to Lowe. *Analysis*, 48, 65–72.
Lewis, D. (1991). *Parts of classes*. Oxford: Blackwell.
Lewis, D. (1994). Humean supervenience debugged. *Mind*, 103, 473–90.
McLaughlin, B. & Bennett, K. (2011) Supervenience. In E.N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*, http://plato.stanford.edu/archives/win2011/entries/supervenience/, accessed June 2, 2014.
Merricks, T. (2001). *Objects and persons*. Oxford: Oxford University Press.
Sider, T. (2001). *Four dimensionalism*. Oxford: Oxford University Press.
Sider, T. (2007). Parthood. *Philosophical Review*, 116, 51–91.
Uzquiano, G. (2014) Mereology and modality. In S. Kleinschmidt (Ed.), *Mereology and location*. Oxford: Oxford University Press.
Varzi, A. (2010). Mereology. In E.N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (Spring 2010 edition)*, http://plato.stanford.edu/archives/spr2010/entries/mereology/, accessed June 2, 2014.
Van Inwagen, P. (1981). The doctrine of arbitrary undetached parts. *Pacific Philosophical Quarterly*, 62, 123–37.
Van Inwagen, P. (1990). *Material beings*. Ithaca, NY: Cornell University Press.
Van Inwagen, P. (1994). Composition as identity. *Philosophical Perspectives*, 8, 207–20.

17

Humean Reductionism about Laws of Nature

NED HALL

17.1 Introduction

Are laws of nature mere patterns in the phenomena (specially *salient*, perhaps – but still, just patterns)? Or are they something more, something that somehow *governs* or *constrains* those phenomena? Disagreement over this issue constitutes the schism in contemporary philosophical work on laws of nature. Here, I will investigate the prospects for an important position that falls under the "mere patterns" approach: what, for reasons that will emerge, I call "Humean reductionism" about laws of nature, a view championed perhaps most prominently by David Lewis. I will review some of the most interesting arguments against this position from the literature, and add some of my own that, I think, are more effective. All the same, "investigate" remains the operative word. The value of the various anti-Humean arguments lies not so much in their power to refute, as in their power to force a sharper and more sophisticated articulation of a Humean position capable of resisting them. So, while I immodestly claim that my own anti-Humean arguments are the best out there, I offer them with only the modest aim of improving our understanding of the schism.

17.2 Preliminaries

I'm going to begin by assuming, without argument, a basic distinction between fundamental laws and the so-called "laws" of the special sciences. I will only be concerned with the former. It is the job of fundamental physics, and fundamental physics alone, to uncover the fundamental laws. These laws describe the way in which complete physical states of the world generate successive physical states – or, if you prefer, the way in which complete physical states are followed and preceded by complete physical states.[1] I take it that the fundamental laws of our world fix a distinction between worlds that are nomologically possible, relative to our own, and worlds that are not.

It will be useful to have an example of some fundamental laws, to use as a test case. Current physics is too messy; so we will rewind about two centuries, and consider Newtonian particle

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

mechanics. That is, we will assume that our world is wholly composed of some finite number of indestructible point-particles, moving in continuous trajectories through absolute space,[2] and each possessed of an unchanging value for mass and charge. We'll assume that these particles move in accordance with Newton's laws of motion and the appropriate force laws. These laws pick out a range of *nomologically possible total histories of particle behavior*.

A handy way to *map* these histories is as follows: we will suppose that the fundamental laws delimit a set of nomologically possible *initial conditions*, each of which specifies the number of particles at the "initial time," along with their masses, charges, positions, and velocities. (It would be enough to assume that the laws delimit a set of possible instantaneous physical states for the world as a whole to be in, at whatever time. But we will stick with the simplifying assumption that there is an initial time.) Second, the laws specify how each such initial condition would *evolve forward* in time. Thus, the fundamental laws "factor" the set of nomologically possible total histories of particle behavior into an initial conditions hypothesis (ICH) and a dynamical hypothesis (DH).

Some comments. First, we will assume, in the case of our Newtonian example, that the dynamical hypothesis is deterministic, so that no two nomologically possible histories perfectly agree on particle behavior (i.e., on their distribution of particle positions, masses, and charges) up to some time without agreeing for all time.[3]

Second, I said that ICH + DH yields the set of nomologically possible histories of particle behavior – not the set of nomologically possible *worlds*. One reason for holding back from the stronger characterization is that, according to most anti-reductionists about laws of nature, more would be needed to specify the nomologically possible worlds. Minimally, one needs to say that they are worlds that not only exhibit nomologically possible histories of particle behavior, but that are also governed by the *same laws* as the actual world. And that is a substantive addition.

Third, it is worth noting that *breadth* or *permissiveness* of the ICH makes for a certain kind of *explanatory strength*. For it is, other things equal, a point in favor of a physical theory that it recognizes a wide range of nomologically possible initial conditions. Compare, for example, Keplerian and Newtonian accounts of the solar system. Granted that the Newtonian account is much more empirically accurate; it is also, from the standpoint of scientific investigation, better in a distinct sense: for it allows us to *answer questions* not merely about how the elements of the solar system did, do, and will behave, but also about how they *would have behaved* under alternative physical conditions. We will come back to the significance of this kind of strength in section 17.6; we will see that the Humean reductionist has a difficult time accounting for its importance in physical theorizing.

17.3 The Elements of Humean Reductionism

It's usual to set up Humean reductionism as a supervenience claim: all facts – and in particular, all facts about the fundamental laws – supervene on the totality of non-modal facts, so that no two worlds differ with respect to their fundamental laws without differing somehow with respect to their non-modal facts. I find supervenience too weak a notion, and unexplained talk of "non-modal" (sometimes: "non-nomic") facts unhelpful. (For example, if facts about laws supervene on non-modal facts, doesn't that mean that they are *themselves* non-modal facts? If so, the supervenience claim is trivial.) Let me suggest a different way of proceeding.

Start with the assumption, widespread among reductionists and anti-reductionists alike, that there is such a thing as the fundamental ontological structure of the world: facts about the world to which all other facts *reduce*.[4] The Humean reductionist needs a doctrine about the nature of this fundamental ontological structure. Lewis was admirably clear on this point. Drawing inspiration from Hume, he took the fundamental ontological structure to consist in the pattern of instantiation

by spacetime points of perfectly natural monadic properties, together with the facts about the spatiotemporal relations among those points. Now, his emphasis on *properties* and *relations* looks a bit clumsy, given that it is *magnitudes* that are the stock in trade of real physics. But we'll let this pass, as the needed amendments are straightforward (see Hall 2012 for discussion). These facts count as "non-modal" precisely because the metaphysical possibilities involving the perfectly natural properties and relations respect a principle of recombination, according to which, for any combinatorially possible way of distributing perfectly natural properties and relations among particulars, there is a metaphysically possible world in which they are so distributed. (Lewis thereby captures the Humean idea that there are no necessary connections between distinct existences.) So an improved Humean reductionism about laws would be this: facts about the laws reduce to facts about the distribution of perfectly natural properties and relations. That's much better.

I am going to take for granted that the Humean reductionist intends this thesis to state a *necessary* truth. There are good reasons for such a stance, though it would take us too far afield to review them. At any rate, nothing in what follows will hinge on this point.

Applied to our Newtonian particle world, finally, the thesis is that the facts about its laws reduce to the facts, over all time, about the positions, masses, and charges of all the particles. The next question is *how*.

17.4 The Best System Account: An Overview

Lewis, citing inspiration from Mill and Ramsey, famously defended a certain sophisticated treatment of laws as mere *regularities*. There are, I think, two distinct guiding ideas behind his account: the one that typically gets official mention, and a second one working behind the scenes. (Warning: I will mostly be ignoring the fascinating issues that arise in adapting Lewis's account to the treatment of probabilistic laws. Though see Hall 2012 for extensive discussion.)

The official idea is that the laws are, or correspond to, true statements that collectively encode, in a highly efficient manner, a large amount of information about the world. The unofficial position is a kind of "ideal observer" view, according to which the fundamental laws are whatever a suitably placed observer, implementing the best scientific standards for judging what laws are, would take them to be. I will consider these two ideas in turn, and then highlight an obvious way in which we might think to reconcile them. (By the end of the chapter, we will see the prospects for reconciliation dim substantially.)

17.4.1 The Official Guiding Idea

Here is a classic statement from Lewis of what we will henceforth call the *Best System account* (BSA):

> Certainly not just any regularity is a law of nature. Some are accidental. So an adequate regularity analysis must be selective. Also, an adequate analysis must be collective. It must treat regularities not one at a time, but rather as candidates to enter into integrated systems. . . . I take a suitable system to be one that has the virtues we aspire to in our own theory-building, and that has them to the greatest extent possible given the way the world is. It must be entirely true; it must be closed under strict implication; it must be as simple in axiomatisation as it can be without sacrificing too much information content; and it must have as much information content as it can have without sacrificing too much simplicity. (Lewis 1983, 41)

Lewis adds that "simplicity" can be achieved illicitly – via a suitably perverse choice of primitive vocabulary – and so requires that "the primitive vocabulary that appears in the axioms refer only to

perfectly natural properties." (I think he should have said: "perfectly natural *magnitudes*" – cf. the point raised in the last section – but I won't fuss over this.)

Focus on the overarching idea. Lewis takes it that there is some *canonical scheme* for representing facts about the world. (Unlike Lewis, I think this canonical scheme should take the form of mathematical models of the kind used in real physics – not sets of sentences in a first-order language. But again, I won't fuss over this.) Any correct representation that makes use of this scheme will have two features:

First, it will have a degree of informativeness, determined purely by which possible worlds the representation rules out. So it automatically follows that if one correct representation rules out more possible worlds than a second (i.e. every world in which the second true is one in which the first is true, but not vice versa), then the first is more informative. There are thus maximally informative representations, made so by being true only in the actual world.

Second, it will have a degree of simplicity, determined by broadly syntactic features of the representation ("broadly," because we want to leave it open whether the canonical scheme of representation is linguistic; see the parenthetical remark above). These two factors of simplicity and informativeness then determine an ordering – presumably, partial – among all the correct representations there are, in terms of how well each one *balances* simplicity and informativeness. Lewis's hope is that the nature of our world will yield a clear winner.

Let's briefly consider how the BSA would apply to our Newtonian particle world. We want to capture as much information as we can about the motions, masses, and charges of particles by exhibiting equations that particle behavior must conform to. If we were allowed to invent any old magnitude we wished, allowing a parameter representing it into our equations, then this would be easy: define a "magnitude" M that is stipulated to have value 1 for a particle exactly if that particle belongs to a world of particles whose behaviors are just what they actually are. But if we require that the parameters that appear in the equations correspond to *perfectly natural* magnitudes – that is, position, mass, and charge – then presumably things aren't so easy. It is not at all implausible that the equations that appear in Newtonian mechanics achieve, among all such candidate equations, far and away the best balance of simplicity and informativeness.

17.4.2 The Unofficial Guiding Idea

So much for the first guiding idea. The second is, in fact, already hinted at in the first of the passages quoted from Lewis. Here it is again: "…I take a suitable system to be one that has the virtues we aspire to in our own theory-building, and that has them to the greatest extent possible given the way the world is." Elsewhere, he is even more explicit: "The standards of simplicity, of strength, and of balance between them are to be those that guide us in assessing the credibility of rival hypotheses as to what the laws are" (Lewis 1986, 123).

To help flesh this idea out more fully, let us imagine someone whom I will call a Limited Oracular Perfect Physicist. What makes our LOPP a perfect physicist is that, given as evidence any information about the world, she is perfectly able to judge what hypotheses about the fundamental physical laws are most strongly supported by that evidence. What makes her oracular is that she has, as evidence, quite a lot of information about the world. But not, of course, *all* information: else her job would be trivially easy. Specifically, we will suppose her evidence consists of all the facts about the distribution of perfectly natural magnitudes. Hence, in our Newtonian particle world: perfect information about the motions, masses, and charges of every particle, together with the further information that the world contains nothing else.

The second guiding idea, then, is that the laws are whatever she says they are.

Some comments. First, and most obviously, the device of introducing a fictional LOPP is dispensable. What the second guiding idea really needs to assume is that there are, implicit in the practice of physics, evidential standards for determining what the fundamental physical laws are that induce a mapping from possible *total bodies of evidence* to something like a probability distribution (or perhaps a family of such distributions) over propositions about the fundamental laws of nature. It is nothing more than a useful heuristic to imagine a creature who holds that mapping in her head. Next – and this is quite important – it is no part of anti-reductionism per se that it must reject the existence of such a mapping. That is, an anti-reductionist can – and should – perfectly agree that there is a fact of the matter about how one's credence over alternative hypotheses about the fundamental laws of nature ought to be distributed, were one fully apprised of the positions, masses, and charges of every particle at every time. (In the case of our well-behaved Newtonian world, the anti-reductionist presumably will say that one ought to be all but certain that the laws are the Newtonian laws.) But the anti-reductionist will view this as a purely *epistemic* fact of the matter. So we can summarize the second guiding idea this way: the Humean reductionist is taking standards that *both* sides endorse – but that his anti-reductionist opponent views solely as epistemic standards – and elevating them to the status of standards constitutive of the laws of nature.

One subtlety needs more comment. It is not legitimate for our Humean reductionist to assume that the epistemic standards implicit in the best practice of physics have a built-in bias toward *his* metaphysical position about laws. In fact, what makes the most dialectical sense is for him to assume the *opposite* (and, for reasons that we will get to later, this assumption is independently plausible). He will then simply need to compensate – with luck, only slightly – for this anti-reductionist bias. Let me explain.

Imagine that we are reductionists, and that we are interviewing our LOPP. She has been informed of the complete history of particle motions, masses, and charges, and has formed in response an opinion about what the laws are, hence which worlds are nomologically possible. We ask her to tell us which worlds these are. She tells us that since, of course, facts about laws do not reduce to non-modal facts, in order to tell us which worlds are nomologically possible she must tell us two things: what a given world's history of particle behavior is, *and* what its laws are. Never mind that, according to her, the second part is easy, since it is nomologically impossible for the laws to be other than they *actually* are. We are, all the same, distressed. For she is not giving us exactly the information we want – indeed, she is giving us "information" (namely, about metaphysically irreducible laws) that we think does not exist. So we rewind, and ask our question more carefully: "Please, oh LOPP, tell us which worlds, described up to but not beyond their non-modal features, are nomologically possible." We need not worry that her answer to *that* question will presuppose anything objectionable. The moral is that our slogan formulation of the second idea went slightly awry. We should have said that the Humean reductionist takes epistemic standards *for determining the nomological possibilities for the non-modal facts* – standards that both he and his opponent endorse as such – and promotes them to the status of standards constitutive of the nomological possibilities for the non-modal facts. (The reductionist then adds, of course, that these are *all* the facts.)

17.4.3 Reconciling the Two Guiding Ideas

So we have two, apparently distinct, guiding ideas. I will close this section by highlighting the obvious way to try to knit them together. Take the second guiding idea – that our implicit scientific standards for judging lawhood are in fact *constitutive* of lawhood – to be the central, non-negotiable one. Take the first idea – that lawhood is a matter of balancing simplicity and informativeness – to be a *substantive proposal about the nature of these implicit standards*. Put another way, "candidate systems" can be

thought of as constituting the range of hypotheses that our LOPP considers. She employs some criteria for ranking them in terms of plausibility, and what the first guiding idea tries to do is to say explicitly what those criteria are.

Later on I will cast doubt on whether this reconciliation will really work. But for now it will be very useful to keep in mind when reviewing certain challenges to the BSA, for it will be important to distinguish those challenges that target the reductionist's specific proposal to analyze laws in terms of simplicity and informativeness, from those that target the underlying idea that the practices implicit in physics for judging lawhood can be used to say what laws *are*.

The next section takes a selective look at two of these challenges. (For a much less selective look, see Hall 2012.) The last of them will require a substantial revision to the BSA, presented in section 17.6. This revision leads in turn, I think, to a serious dilemma for the Humean reductionist. I will close in 17.7 with some speculative remarks about how the Humean might best overcome this dilemma.

17.5 A Menu of Challenges to the BSA

The BSA has been challenged in a wide variety of ways: It violates certain (non-negotiable?) intuitions we have about laws. It makes a hash out of the connection between laws on the one hand, and counterfactuals, explanation, and induction on the other. It fails to yield an acceptable account of nomological possibility. It gives rise to insuperable skeptical problems. It cannot makes sense of what is, intuitively, an essential connection between possession of a fundamental magnitude such as *mass* and the exercise of certain causal powers. It runs afoul of a key principle connecting credence with objective chance.

Set all these challenges aside; the reductionist has resources more than adequate to meet them, or at least keep them at bay. (See Hall 2012 for details.) Focus instead on two more. The first can be met – but in a way that is, I think, highly instructive. The second can also be met – but in a way that leads directly to the more serious problem alluded to above.

17.5.1 The Problem of Ratbag Idealism

The first challenge to consider concerns the pedigree of the standards of informativeness and especially simplicity. As Lewis puts it:

> The worst problem about the best-system analysis is that when we ask where the standards of simplicity and strength and balance come from, the answer may seem to be that they come from us. Now, some ratbag idealist might say that if we don't like the misfortunes that the laws of nature visit upon us, we can change the laws – in fact, we can make them always have been different – just by changing the way we think! (Talk about the power of positive thinking.) It would be very bad if my analysis endorsed such lunacy. (Lewis 1994, 479)

An effective reply beckons.[5] For it is at this point that the second guiding idea – that the standards for judging candidate systems should be the very standards that an anti-reductionist endorses as appropriate *epistemic* standards for figuring out what the laws are – comes into its own.

Perhaps it was a mistake for the BSA to implicitly assume that the *simplicity* of a theory is one of the things we focus on as a guide to its truth. If so, the problem is *not* with the core idea behind the BSA but with its particular implementation. So suppose otherwise. Suppose, that is, that *everyone*

should agree that one of the considerations our LOPP should use in deciding which of the various law-hypotheses is correct is its simplicity. The anti-realist insistently reminds us that these standards are merely *epistemic* standards, standards such that we ought, as a normative epistemic matter, to believe that hypothesis that scores highest with respect to them. But then the problem of the ratbag idealist is very much *her* problem. In fact, it is even *more* her problem. For consider: If it is somehow "up to us," in an objectionable sense, what counts as simple, then, by the anti-reductionist's lights, central facts of normative epistemology are *also* up to us: facts about what one ought to believe about the laws, in the face of empirical evidence. To parrot Lewis: "Now, some ratbag idealist might say that if we don't like the misfortunes that our epistemological principles have visited upon us, we can change these principles – in fact, we can make them always have been different – just by changing the way we think!"

That, I think, is a deeper lunacy than the kind Lewis seeks to evade. By contrast, the reductionist position seems perfectly sane, at least by its own metaphysical lights. And here, just to be clear, I am parting company with Lewis. I think he was deeply mistaken to think that there was *any danger whatsoever* posed by the problem of ratbag idealism. It will be instructive to see why.

Reflect, for a moment, on the implications of the basic metaphysical outlook adopted by the reductionist. It helps to keep firmly in mind that he thinks that, for any world, *all there is* to that world is a distribution throughout space and time of various perfectly natural magnitudes. For example, all there is to our Newtonian particle world are some particles moving around, with masses and charges. That's it. It is emphatically *not* that the facts about these particles serve as clues to something "behind the scenes" that is directing their behavior. That is quite the wrong way to think about it. In fact, a much better way to think about the status of laws, given such a background metaphysics, is *pragmatically*.

An example will help clarify what I have in mind. Suppose you are a biologist. You are no vitalist. You face the question what, given your anti-vitalist commitments, should count as "life." The first thing you should do is to scrap that question, and replace it with this one: What, given your anti-vitalist commitments, is the best – that is, theoretically most useful – way to draw a distinction between things in the world that might roughly correspond to our ordinary living/non-living distinction? You might find a number of reasonable candidates. You might find that the choice between them should be made entirely on grounds of theoretical utility – indeed, that for some purposes one way of drawing the distinction is useful, for other purposes another. So too, I claim, with laws – at least, *given* a reductionist metaphysics. We have an array of non-modal facts about a world – as it might be, facts about how many particles there are, what their masses and charges are, and how they are moving. Science, presumably, is in the business of investigating *that* stuff. It is not in the business of investigating any further stuff that lies behind the scenes, for this "scenery" just *is* reality. So do not ask what, given such a metaphysical outlook, laws are. Ask instead how, given such a metaphysical outlook, one might usefully draw a distinction between certain facts about the world that are in some sense *distinctively appropriate targets for scientific inquiry* (DATSIs), and other facts that are less interesting and central.

For now, I simply wish to note that there is plenty of room for the view that it is in part facts about us – *idiosyncratically* about us, and our peculiar human psychologies – that play a role in determining the most appropriate way for us to structure our investigation of the world. How could the details of our peculiar human situation *not* be relevant to this matter? And if so, then *of course* it should be partly "up to us" – that is, partly contingent on our particular needs and limitations as inquirers – what counts as a DATSI.

Here, the contrast with the anti-reductionist could not be sharper. On her view, the laws governing a world demarcate a range of facts about the world profoundly, fundamentally distinct from any other kinds of facts, a range of facts with deep and interesting metaphysical connections to the realm of

non-modal facts. And given that they are empirically accessible at all, only the most remarkably incurious person could fail to think that scientific inquiry ought to be structured in part with the aim of uncovering them. By contrast, a reductionist cannot think of laws as having some such natural epistemic magnetism to them. To put the point pithily, for the anti-reductionist, the nature of laws *automatically* makes them DATSIs; whereas for the reductionist, it is exactly the reverse. On his view, all it could come to for a claim to achieve the status of law is that, for some other and *independent* reason, it is a DATSI.

Verdict: It would have been smarter for the anti-reductionist to keep quiet about the ratbag idealist. But taking stock of the problems this creature does and does not pose allows us to sharpen our understanding of the differences between reductionism and anti-reductionism. Section 17.6 will take advantage of this improved understanding.

17.5.2 The Information That Counts

We come now to a problem that, I think, requires a substantial amendment to the BSA. Suppose that the BSA, when applied to our Newtonian particle world, succeeds to this extent: the best system – the one, remember, that optimally combines simplicity and informativeness – includes the Newtonian dynamical laws. Suppose it also includes a broad constraint on initial conditions, so that the combined result is that it counts a world as nomologically possible only if that world contains a finite number of particles, each possessing some value of mass and charge, and moving on trajectories that perfectly conform with the Newtonian dynamical principles. So far, so good. The problem is that there is every reason to think that this "only if" cannot be strengthened to an "if and only if"; that is, that the best system for our Newtonian world will exclude *too many* worlds that ought to count as nomologically possible.

A quick way to see this is to notice that we get a vast increase in informativeness, at only a small cost in simplicity, if we include in a candidate system a statement that says exactly how many particles there are. Only *small* cost in simplicity? Well, the cost might be middling, if the number of particles is a large number that is not mathematically identifiable in a simple way. But suppose it is. Suppose it turns out that there are exactly 2^{64} particles in our world. Then a statement to that effect buys a lot of informativeness at a negligible cost in simplicity. And it seems quite wrong to think that, *for that reason*, it ought to count as nomologically necessary that there are 2^{64} particles. In the first place, particle number should turn out to be a nomologically contingent feature of the world. In the second place, even if you have somehow convinced yourself that there can be good grounds for considering particle number to be a nomologically necessary feature of the world, surely those grounds cannot be that this number is easily mathematically describable.

At any rate, there are more pressing versions of this problem. Suppose, for example, that there is some moment of time such that the complete physical state of the world at that time happens to have a very simple to describe form: there is some relatively simple, compact way to say exactly what that state is. Let this state be S. Then, if a candidate system includes the Newtonian dynamical principles, it buys an *enormous* increase in informativeness by adding a statement to the effect that at some time, the complete physical state of the world is S. For doing so will shrink the set of nomological possibilities down to *one*. (Here I am taking advantage of the fact that Newtonian dynamics are two-way deterministic.) But that is a disaster, for remember that it is a non-negotiable desideratum on an account of laws that it yield a non-trivial distinction between what is nomologically possible and what is not.

Finally, it is a disaster that is unavoidable, essentially because it is guaranteed that our world exhibits a simple to describe state. For consider that the state of the world at any time can be coded

up, in a very simple way, by a single real number: just take all the coordinates, masses, and charges of all the particles, expressed in decimal notation, and interleave the digits. Suppose we include this number in a candidate system; then, once again, we get an increase in informativeness that shrinks the set of nomological possibilities down to one. Call this the problem of *the phony fundamental constant*.

"But," comes the response, "don't we get this increase in informativeness at the price of a huge increase in complexity? For even though the way in which this real number codes up complete information about a physical state is simple, the number *itself* is very unlikely to be at all simple. That is, there will be no mathematically compact way to specify its value."

I do not think there is any refuge in this response. Remember that it is part of the practice of physics to include in its theories *fundamental constants*. But it is *no* part of the practice of physics to insist that those constants have mathematically tractable values. Pick any mathematical standard of complexity of number you like, and a physicist will say that for all she knows, the fundamental constants might be *that* complex. (And do not think that we can buy mathematical simplicity by a suitable choice of units. The quickest – though not the only – way to see that this won't work is to remember that some of the constants appearing in our current physical theories are unitless.) Yet it is nomologically necessary that the constants have the values they do.

More to the point, there is something silly in all of these defenses. For they miss what is really going wrong in these problem cases, which is that a candidate system is scoring points for the *wrong kind of informativeness*. Recall how, in section 17.2, we factored our Newtonian specification of nomological possibilities into two parts: there is an initial conditions hypothesis, that specifies what initial conditions are nomologically possible; and there is a dynamical hypothesis, that specifies how each such initial condition would evolve. The problems we are considering have the common feature that a candidate system wins merely because it narrows down the range of possibilities allowed by the ICH. That, I claim, is an increase in informativeness that ought not to count in favor of a candidate system that exhibits it.

Stronger: it ought to count *against* a candidate system that exhibits it. (This idea is not new: compare Earman's discussion in his 1984.) Part of the reason involves an appeal to intuition: facts about initial conditions seem, intuitively, to be the sorts of things that a good physical theory will treat as contingent. But a better reason focuses on the central explanatory practices of physics, which make essential use of counterfactuals concerning worlds with different initial conditions, but governed by the same dynamics as our world. What, for example, explains the fact that the planetary orbits deviate from perfectly elliptical orbits? The presence of other planets, and corresponding perturbation by interplanetary gravitational forces: for *if all other planets were absent*, then a given planet would orbit the sun in a nearly perfect ellipse. More generally, I claim, the ability to provide sharp and determinate truth conditions for a wide range of counterfactuals is precisely what lends a good physical theory its explanatory power. So, by unduly restricting the ICH, a candidate system will hamstring its ability to evaluate counterfactuals, and thereby fatally undermine its explanatory potential. So it was a serious mistake for the BSA to incorporate a standard of informativeness according to which a candidate system automatically improves its score by narrowing the range of worlds it counts as nomologically possible. *This is not the kind of informativeness that counts*.

Verdict: The official formulation of the BSA needs a substantial revision, in the form of an account of informativeness that is more sensitive to the actual practice of physics. The next section provides just such an account, and produces an elegantly amended BSA.

But victory for the reductionist will be short-lived, for he will be hard-pressed to explain why, by his metaphysical lights, the amended BSA should be preferable to the original version.

17.6 A Solution and a Problem

17.6.1 Solving the Problem of Too Much Information

If the diagnosis just provided is correct, then it is easy to see how to produce an improved version of the BSA. Here is a sketch:

Let a candidate system have two parts: an ICH, and a DH. A system is better to the extent that each of these parts can be specified in a simple way. But in the case of the ICH, a system is better to the extent that this hypothesis is *un*informative – that is, to the extent that it admits *more* initial conditions as nomologically possible. Where we want a candidate system to be informative is, instead, solely in its dynamical hypothesis. And, with respect to the informativeness of the DH, we can say something much more illuminating about what we mean by "informative" than Lewis did. That is because there is a clear, non-arbitrary *ideal* of informativeness for a dynamical hypothesis to aim for: determinism, in both temporal directions.

Next, there are fairly clear and non-arbitrary ways of falling short of the ideal of determinism. For example, a DH might be deterministic only towards the future. Or, it might be deterministic for all but a very small measure of initial conditions. Or it might be indeterministic only in a "punctuated" sense, so that only when certain specific and localized kinds of interactions take place does the DH fail to describe their outcomes in a deterministic fashion. (Our chancy Newtonian world was like that.) And, even where the DH falls short of determinism, it might succeed in laying down precise probabilities. It is, I think, a very interesting question how to develop a precise catalog of the varieties of less-than-perfect informativeness that a DH might exhibit. I will not pursue that question further here. Still, it is quite striking how, by shifting attention from the candidate system as a whole to just one component of it – its dynamical hypothesis – we have *already* arrived at a vastly more interesting characterization of "informativeness" then the utterly bland characterization routinely trotted out in the literature defending the BSA, that says only that informativeness is somehow a matter of narrowing down possibilities. It's not, and, even if it were, it would remain an embarrassment that the literature has been unable to say anything more illuminating.

Why is two-way determinism the gold standard of informativeness for a DH? It is not *just* that this kind of informativeness is pretty clearly what physicists (many of them, anyway) would say that they aim for in their theories. They have, in addition, an excellent reason for setting their sights on such theories, since meeting this standard directly serves the theoretical purposes emphasized in section 17.2: a theory that does so can, as a result, answer with great precision questions both about what would have happened, and about what would *have* to have happened, had conditions in the world been different in some specified respect. That is one extremely important way for a theory to increase its explanatory power. The other way, of course, is for the theory to incorporate a maximally permissive ICH.

These observations suggest an amendment to the official guiding idea behind the BSA: what candidate systems aim for is not an optimal combination of simplicity and informativeness, but rather an optimal combination of simplicity and *explanatory power* – where explanatory power in turn is a function of the uninformativeness of the system's ICH together with the informativeness of its DH. As before, the nomologically possible worlds will be exactly those that conform to the candidate system that wins out in this competition.[6]

17.6.2 A Dilemma for Humean Reductionism

Our new, improved BSA immediately and effectively deals with the challenges raised in 17.5.2. What's more, it *seems* to do so in a perfectly well-motivated way: for remember that the intention of

the BSA is to incorporate, as the standards for judging candidate systems, standards implicit in the best practice of physics. And those standards place a premium on physical theories that say a lot about what would have happened, under alternative possible conditions – thereby placing a premium on keeping one's ICH *un*informative.

But now we come to a problem. If we pay close attention to the background metaphysical commitments that underlie the BSA, we can see that it faces a curious but difficult dilemma. Building the ICH/DH distinction into the BSA seems to be necessary, in order to prevent Humean reductionism from being too violently at odds with the actual practice of physics, and in particular with its aim of constructing theories that are maximally explanatory powerful. For example, if physicists treat particle number in a Newtonian world as a nomologically contingent feature of that world, then, at the very least, someone who disagrees with them on purely philosophical grounds is in a highly uncomfortable position. But, however successful the new version of the BSA is at dodging potential counterexamples – and in particular, in securing agreement with the physicists about which initial conditions should be treated as nomologically possible – it is, by the Humean's lights, *entirely unmotivated*.

To see why, we should return to the conception of "law" emphasized in 17.5.1: some claim about the world counts as a law only to the extent that it is, on grounds *independent* of its nomological status, a distinctively appropriate target for scientific inquiry. Now, for the reductionist there is one obvious way to pick out a special set of claims: they will be claims about the kind and distribution of those perfectly natural magnitudes that characterize our world. Such claims have a distinguished metaphysical status, insofar as all *other* facts about the world reduce to them. But, just as obviously, that is not *enough* of a distinction: one wants some *further* way of picking out, among all the true claims about the distribution of perfectly natural magnitudes, an elite subset that qualifies as DATSIs.

An anti-reductionist can say that certain claims about the fundamental, non-modal facts deserve special attention *because they hold as a matter of law*. And so it makes perfect sense to structure inquiry in science – and especially in physics – in such a way that it has a chance of uncovering these laws. But for a Humean reductionist, *all there is*, fundamentally, is an assortment of non-modal facts – facts that are all perfectly on a par, metaphysically speaking. Scientific inquiry is, in the first instance, merely directed at finding out about *them*. It would seem, then, that any further structure to this inquiry can be imposed only by the need to find out about those facts *efficiently*. Helen Beebee has provided an especially eloquent articulation of this idea:

> So the idea is something like this. Suppose God wanted us to learn all the facts there are to be learned. (The Ramsey-Lewis view is not an epistemological thesis but I'm putting it this way for the sake of the story.) He decides to give us a book – God's Big Book of Facts – so that we might come to learn its contents and thereby learn every particular matter of fact there is. As a first draft, God just lists all the particular matters of fact there are. But the first draft turns out to be an impossibly long and unwieldy manuscript, and very hard to make any sense of – it's just a long list of everything that's ever happened and will ever happen. We couldn't even come close to learning a big list of independent facts like that. Luckily, however (or so we hope), God has a way of making the list rather more comprehensible to our feeble, finite minds: he can axiomatize the list. That is, he can write down some universal generalizations with the help of which we can derive some elements of the list from others. This will have the benefit of making God's Big Book of Facts a good deal shorter and also a good deal easier to get our rather limited brains around. (Beebee 2000, 547)

It makes perfect sense, given a reductionist metaphysics, that scientific inquiry strives to figure out exactly which "axioms" God has written down: for doing so directly serves what can be the only *ultimate* aim of inquiry, which is to figure out as much of the non-modal truth about the world as possible (there being no other truth to be had). But that seems to point us, alas, directly to Lewis's

original formulation of the BSA, with its premium on the kind of informativeness that consists merely in the narrowing down of possibilities. He suggested, in one of the passages we quoted, that the standards of simplicity and informativeness are "virtues that we aspire to in our own theory building." If I am right about the proper diagnosis of the problems raised in 17.5.2, this suggestion is mistaken. While simplicity may be such a virtue, it is precisely *uninformativeness* that we aspire to in the part of our theory that tells us what initial conditions are possible. Why should this be a virtue if, in the first instance, all that scientific inquiry consists in is an investigation of what the non-modal facts are about our world?

So the Humean reductionist seems to face the following dilemma: On the one hand, he can avoid conflict with the practice of physics by structuring the BSA around the ICH/DH distinction in the way described above; but it is not clearly intelligible why, by his lights, the resulting standards are the right ones to use in choosing a best system. On the other hand, he can choose standards that make perfect sense, given his metaphysical commitments; but doing so appears to lead straight to the original version of the BSA, a theory that from the standpoint of the actual practice of physics is insane. This choice between a guilty intellectual conscience and insane revisionism is not a happy one.

Consider, by contrast, what metaphysical viewpoint would motivate the importance of the ICH/DH distinction, and in particular an adherence to standards of theory choice that looks for an uninformative ICH, coupled with an informative DH. It is not far to seek: it is the view that information about the counterfactual structure of the world is *sui generis*, and not merely disguised information about its non-modal structure. If that is what you think, then *of course* you will want to build into your theorizing about the world techniques for accurately mapping this *sui generis* counterfactual structure. And that is exactly what the distinction between initial conditions and dynamics accomplishes. Notice, finally, that it is precisely a commitment to a *sui generis* counterfactual structure that unites *all* the disparate varieties of anti-reductionism. So the central feature of physical theorizing that we have focused on cleanly divides the Humean reductionist from his philosophical opponents. His position is not only philosophically uncomfortable, but lonely.

17.7 Conclusion

As I noted at the outset, I think that the problem I have raised – challenging though I hope it is – is really an occasion for the Humean reductionist to sharpen his position still further. I am not sure what the best way is to do this, but here are some preliminary ideas (prompted mainly by conversations with Barry Loewer, one of the most thoughtful reductionists currently writing on the topic).

Focus just on the problem posed by the phony fundamental constant, the one that codes up complete information about the physical state of our particle world at some time. I suggested that the proper way to deal with this problem is to build the ICH/DH distinction into the BSA – but that this suggestion is really a Trojan horse, since the resulting account is one that the reductionist ought to reject, given his metaphysics.

Loewer has suggested (in conversation) a very different response. He observes that one serious defect in a system that incorporates the phony fundamental constant is that it codes up information about the state of the world in a form that is almost completely *useless* to us ordinary humans. To dramatize, suppose an oracle starts to tell you the value of this constant, a few digits at a time. After she has told you a billion digits, you still have something that is virtually useless to you – even if you have unlimited computational power at your disposal. For this information allows you to make almost *no* new predictions. So perhaps the right lesson for the Humean reductionist is to return to Lewis's original formulation, but insist that what is wanted from a candidate system is not merely

information that narrows down possibilities, but information that is, in some suitable sense – and taking into account relevant facts about the condition that we humans find ourselves in – *practically useful*.

One way to spell this idea out focuses on Beebee's remarks about our "rather limited brains." We might reasonably demand that candidate systems pay heed to our feebleness, by representing the world in a way that not only optimally balances simplicity and informativeness, but is such that *approximations to it* – which, given our limitations, may be all that we can ever hope to achieve – are *also* highly informative. Then the real problem with the phony fundamental constant has nothing to do with whether its actual value is mathematically identifiable in a simple way. Even if it *were*, the system that incorporated this value would fail the test we are considering: other, "nearby" systems, that got the value of the phony constant slightly wrong, would do so at the price of losing almost all of the extra informativeness the correct system attains by means of the constant.

What I have just sketched seems to me a promising approach for the reductionist to take, although I have two lingering doubts. First, it seems to me easy to come up with examples of statements about the world that ought to be viewed as nomologically contingent, but that nevertheless contain quite a lot of information of great practical utility to us. I imagine, for example, that cosmologists would very much like to know the total mass of the universe – and that even approximate information about this mass would be enormously predictively and explanatorily valuable. But would their ability to squeeze useful predictive and explanatory information out of this knowledge do anything to show that facts about the total mass should be viewed as nomologically necessary? To me, it seems not – although I freely confess that I have nothing but intuition to go on here, and that these intuitions may have been corrupted by evil anti-reductionist influences.

Second, it seems clear that, however the strategy just sketched is developed, it will still end up making the importance that the practice of physics places on the ICH/DH distinction seemed misguided; it remains to be seen whether the reductionist can supply a plausible "error theory" of this practice, or whether instead his position ends up looking too revisionist for comfort.

Might the reductionist reasonably dispute my claims about the importance of the ICH/DH distinction in actual scientific practice? I very much doubt it, and will offer two hypothetical scenarios to help align your judgments with mine.

For the first, imagine a happy vision of our world 100 years hence, suffering neither war nor poverty, but blessed instead with a fervent commitment on the part of humanity to the advancement of physics. Vast resources have gone into the construction of the SuperUltraDuper-Collider, so large that it must ring the equator. But it's worth it: for the SUD-C will allow our scientists to slam particles together at energies never before witnessed in the history of the universe, and by their behavior resolve the key remaining uncertainties as to what the fundamental laws are that govern that universe. Two theories – call them "Safe" and "Sorry" – remain live options. The mathematics involved in each is formidable, so much so that as preparations are under way for the Crucial Experiment, theoreticians are feverishly performing calculations, so that they will know how to properly interpret the data.

Billy and Suzy, two especially gifted mathematical physicists who work as a team, finish their calculations, horrified at what they have discovered. They run to stop the experiment. It's almost too late – they realize they won't reach the control room in time – but they successfully distract the scientists who are about to push the button by throwing rocks through their window.[7] And a good thing: for they have found that if Safe is correct, then the experiment, if conducted, will go off without a hitch, and indeed produce data that will make it all but certain that Safe is correct (to within the usual limits of inductive uncertainty, of course). But if Sorry is correct, then the conditions produced in the experiment will set off a second Big Bang, with the unfortunate result that no one will ever

know that Sorry is correct. The scientific community decides, with regret, to forgo the experiment, and the SUD-C is dismantled. After all – wait for it! – better Safe than Sorry.

Jokes (such as they are) aside, there is something quite interesting and noteworthy about the way our imagined scientists reason about the practical problem they face. Should they conduct the experiment? "Well," they will say, "it *might* be that were we to conduct it, everything would turn out just fine. But it *also* might be that the world would be destroyed. The second possibility is, given our evidence, sufficiently likely that we cannot take the risk." Observe that they are *presupposing* that one of these counterfactuals is true (*determinately* true, for what it's worth), and that presupposition appears to play an indispensable role in their practical deliberations. I submit that *this* part of the story isn't fiction: what I have described is exactly how any good scientist – or good thinker, uninfluenced by Humean ideology – *would* deliberate, were they in such a situation.

Our scientists would deliberate in this matter even if they were convinced that, without their experiment, the world would *never* exhibit behavior, detectable by us or otherwise, that would settle (even to a reasonably high probability) which of Safe and Sorry was correct. (And they might reasonably be so convinced, given the unique nature of the conditions they are aiming to create.) But if this conviction were correct, then by the Humean's lights there would simply be no fact of the matter about what the relevant fundamental laws were – and so, no fact of the matter about what would happen, were the experiment to take place. That's intuitively odd: were we to act in a certain way, our doing so would make it the case that there was some determinate fact about what would happen, were we to act in that way. But more to the point, *it's not the way any sensible physicist would think about this situation*. That is because they are taking for granted that the world possesses – at least with respect to this experiment – a definite counterfactual structure, of the kind that the ICH/DH distinction is designed to chart.

A second way to use this example will pump intuitions even more effectively, I think. (And remember: all I am trying to call forth are judgments, on the part of those of you with some familiarity with scientific practice, about metaphysical presuppositions implicit in that practice.) Suppose, this time, that Sorry isn't so sorry: if that theory is correct, then the experiment will simply produce data that says so, with no ill side-effects. But suppose, again, that without the experiment, the world will fail to serve up behavior that settles which of Safe and Sorry is correct. As noted, according to the BSA a *metaphysical* conclusion follows: for it to be fully determinate – in a *non-epistemic* sense – what the laws are, the experiment must take place (or one like it, later). Without it, there will be, we can suppose, a fair amount of determinateness in the laws; but a crucial piece will be missing.

What is mystifying is how it could possibly be important, by a *Humean's* lights, to carry out the experiment. Here you are, the Scientist in Charge in this world. You are thoroughly convinced of the rightness of Humean reductionism. You contemplate whether to order the construction of the SUD-C. It's *expensive*. What will all that money buy you? Not any sort of a *discovery*: even by your own Humean lights, it's not that you would be finding out about some basic aspect of physical reality *of which you would otherwise be ignorant*. On the contrary: if you forego this experiment, then you can be forever confident that you know *all the facts about the fundamental laws that there are to know* – namely, that they are indeterminate as between the Safe-laws and the Sorry-laws. So why spend the money? If you're rational, you won't – because you recognize *no reason whatsoever* in favor of doing so. And in that you are, I reckon, at odds with just about every one of your colleagues.

Suggestive as they are, examples like this succeed, at most, in showing how deep run anti-reductionist sympathies in the practice of physics. They don't refute Humean reductionism. But they increase the pressure on it to, as it were, explain itself. And so it is clear what the debate needs next: a version of reductionism that gets much more serious than Lewis ever did both about what, *precisely*, the standards are for judging candidate systems; why – given the reductionist's metaphysical commitments – those *ought* to be the standards; and why the resulting conception of laws of nature makes

good sense out of, or perhaps allows us to correct errors in, the actual practice of physics. (To be fair, the debate also needs versions of anti-reductionism that present transparently clear conceptions of what the extra features of the world are that go beyond its distribution of perfectly natural magnitudes, and of how empirical evidence can be a reliable guide to these features.) And it should help a great deal to attend to the distinctive sort of work that a concept of "law of nature" performs in actual scientific practice. That is a methodological approach that sensibly conceives of inquiry into metaphysics as continuous with scientific inquiry: not because it is the scientists who really ought to be settling our metaphysical disputes for us (hah!), but rather because an understanding of the fundamental ontological structure of our world will be achieved, if at all, only by paying close attention to the structure of our scientific theories, and most importantly to the reasons why they should or should not have the structure they do.

Acknowledgments

Thanks to Barry Loewer and Tim Maudlin, especially, for help in thinking through the issues in this chapter. A much more comprehensive treatment can be found in Hall 2012.

Notes

1 What is a complete physical state, in general relativity? It will do to let it be the complete physical state of one space-like hypersurface, treating the laws now as constraining how the rest of spacetime can be, compatible with a specific state for some such hypersurface.

2 And no, of course we don't need to assume that space is absolute: we could speak of trajectories in a Galilean spacetime instead. (For a lovely discussion, see Maudlin forthcoming.) But, here, we're keeping things simple, and easy to grasp.

3 Note that the assumption of determinism is in fact problematic, although not for reasons that will matter here. Xia 1992, for example, describes a fiendishly clever five-particle Newtonian world in which the five particles, solely as a result of gravitational forces, accelerate to infinity in finite time. Such a world is clearly not backwards deterministic, and time-reversibility of the dynamics shows that it is not forwards deterministic, either. So it is actually a somewhat delicate matter what extra constraints to impose on the ICH in order to get determinism, assuming the DH is held fixed.

4 If you are at a loss as to what "reduce to" means, and find insufficiently illuminating the observation that it is a logically stronger relation than supervenience, then the following observations may help. There is a characteristic kind of question we ask when doing applied metaphysics: "What is it for such-and-such a fact to obtain?" For example, we might ask, "What is it for something to persist through time?" A four-dimensionalist will give one answer (it is for it to be composed of momentary time-slices, such that for each of the times in question, one of the constituent time-slices exists at that time); a three-dimensionalist will give a different answer (it is for it to be "wholly present" at each moment at which it exists, whatever exactly that means). Or we might ask, "What is it for someone to act freely?" A typical compatibilist will say that it is for her actions to be caused in the right sort of way by internal psychological states (never mind whether those psychological states themselves have fully sufficient causes in the remote past); a typical incompatibilist will say that it is for her to exercise a kind of metaphysically primitive causal agency. I take it that a correct answer to a question of the form, "What is it for such-and-such a fact to obtain?" must display other facts to which the fact in question *reduces*. One can then repeat the question, targeting these other facts. (What is it for a time-slice to exist at a time? Answer: it is for the contemporaneous spacetime points that constitute it to exist.) The fundamental ontological structure of the world is what one arrives at when these questions stop. To say that all other facts reduce to facts about this structure is to say that it is facts about this structure, and only about this structure, that figure in the ultimate answers to "what is it for" questions.

5 Much more effective, I think, than Lewis's – which is to insist that our standards are objective *enough*, and to maintain a pious hope that "If nature is kind, the best system will be *robustly* best – so far ahead of its rivals that it will come out first under any standards of simplicity and strength and balance." (Lewis, 1994, 233)
6 And as before, the desiderata we are seeking to maximize will be in some tension with each other. Having resolved candidate systems into two distinguishable parts – the ICH and the DH – we can say quite a bit more than defenders of the BSA usually do about the nature of this tension. See Hall 2012 for some preliminary remarks.
7 Suzy's gets there first, breaking the window. But Billy's would have done. So she magnanimously agrees to share the credit with him. Philosophers – who, distressingly, have still not figured out the nature of causation – engage in endless disputes about whether she should have done so.

References

Beebee, H. (2000). The non-governing conception of laws of nature. *Philosophy and Phenomenological Research*, 61, 571–93.
Earman, J. (1984). Laws of nature: the empiricist challenge. In R.J. Bogdan (Ed.), *D.M. Armstrong*, pp. 191–223. Boston: Reidel.
Hall, N. (2012). Humean reductionism about laws of nature: director's cut. Unpublished manuscript, a much more comprehensive version of the present chapter.
Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, 61, 343–77.
Lewis, D. (1986). *Philosophical papers*, Vol. II. Oxford: Oxford University Press.
Lewis, D. (1994). Humean supervenience debugged. *Mind*, 103, 473–90.
Maudlin, T. (forthcoming). *Philosophy of physics*. Princeton, NJ: Princeton University Press.
Xia, Z. (1992). The existence of noncollision singularities in Newtonian systems. *Annals of Mathematics*, 135, 411–68.

Further Reading

Albert, D. (2000). *Time and chance*. Cambridge, MA: Harvard University Press.
Arntzenius, F. & Hall, N. (2003). On what we know about chance. *British Journal for the Philosophy of Science*, 54, 171–9.
Cartwright, N. (1999). *The dappled world*. Oxford: Oxford University Press.
Chalmers, D. (2002). Does conceivability entail possibility? In T. Gendler & J. Hawthorne (Eds.), *Conceivability and possibility*, pp. 145–200. Oxford: Oxford University Press.
Elga, A. (2004). Infinitesimal chances and the laws of nature. *Australasian Journal of Philosophy*, 82, 67–76.
Hall, N. (2004). Two mistakes about credence and chance. *Australasian Journal of Philosophy*, 82, 93–111.
Hall, N. (2005). Causation and ceteris paribus laws. *Harvard Review of Philosophy*, 13, 80–99.
Ismael, J. (1996). What chances could not be. *British Journal for the Philosophy of Science*, 47, 79–91.
Lewis, D. (1980). A subjectivist's guide to objective chance. Reprinted in Lewis (1986), pp. 83–113.
Loewer, B. (1996). Humean supervenience. *Philosophical Topics*, 24, 101–27.
Loewer, B. (2001). Determinism and chance. *Studies in History and Philosophy of Modern Physics*, 32B, 609–20.
Maudlin, T. (2007a). *The metaphysics within physics*. Oxford: Oxford University Press.
Schaffer, J. (2003). Principled chances. *British Journal for the Philosophy of Science*, 54, 27–41.
Shoemaker, S. (1980). Causality and properties. In P. van Inwagen (Ed.), *Time and cause*, pp. 109–35. Dordrecht: Reidel.
Tooley, M. (1977). The nature of laws. *Canadian Journal of Philosophy*, 7, 667–98.
Vranas, P. (2002). Who's afraid of undermining? Why the principal principle might not contradict Humean supervenience. *Erkenntnis*, 57, 151–74.

18

Why Lewisians Should Love Deterministic Chance

RACHAEL BRIGGS

18.1 Chance and Determinism

Some events are determined: history and the laws of nature jointly entail that they will occur. And some events are chancy: they have objective chances in the open interval (0, 1). We might ask whether there events that are *at once* determined and chancy – that is, whether the world contains *deterministic chance*.

Lewis (1980b, 118) claimed that deterministic chance was impossible: "To the question of how chance can be reconciled with determinism . . . my answer is: *it can't be done*." But deterministic chance seems ubiquitous in casinos, in statistical mechanics, and in evolutionary theory. It would be best for Lewis's metaphysics if, in spite of what he says, we could reconcile his core views with deterministic chance. I will argue that we can.

Before I begin, I will briefly rebut two Lewisian objections to deterministic chance that I take to be non-starters. The first is that our world indeterministic at the quantum level, and this lower-level indeterminism translates to indeterminism at higher levels.

> There is chance enough in the processes by which the coin leaves the fingers; in the processes whereby it bounces off air molecules and sends them recoiling off, perhaps to knock other molecules into its path; in the process whereby the coin does or doesn't stretch a bit as it spins, thereby effecting its moment of inertia; and in the processes whereby it settles down after first touching the table. In ever so many minute ways, what happens to the coin is a matter of chance. (Lewis 1980b, 118–19)

This argument is flawed for three reasons.

First, the microphysical world may turn out to be deterministic, as Bohmians claim. Second, even if quantum mechanics turns out indeterministic, there is no guarantee that the indeterminacies will percolate up to the macroscopic level. And third, even if the indeterminacies do percolate up, they are unlikely to percolate up in the right way. We typically think not just that coin tosses are chancy, but that two tosses of the same coin yield the same chance of heads. This assumption is crucial to

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

the practice of ascribing chances to events, both for the purpose of predicting outcomes based on a theory (it would be bizarre to post different betting odds for two tosses of the same coin) and for the purpose of confirming a theory based on outcomes (when determining the coin's bias, we treat all tosses the same way). Lewis's proposal gives us no reason to think – in fact, gives us good reasons to doubt – that repetitions of what we ordinary consider the same experimental setup will yield the same chances. After all, every coin toss involves a completely different arrangement of air molecules.

The second non-starter is the complaint that deterministic chance is "counterfeit chance" (see Lewis 1980b, 121) – it resembles real chance, but is really something else. While this may be a suitable fallback position for someone with independent reason to reject deterministic chance, it is not itself a reason to reject deterministic chance. In order to assess whether deterministic chance is genuine chance, we must ascertain what genuine chance is. It is to that task that I now turn.

18.2 Lewis's Theory of Chance

In order to understand Lewis's theory of chance, we must contextualize it within his broader philosophical methodology. Lewis (1983, x) describes this methodology as a search for reflective equilibrium between our theoretical commitments and our common sense beliefs.

> Our intuitions are simply opinions; our philosophical theories are the same . . . they are all opinions, and a reasonable goal for a philosopher is to bring them into equilibrium . . . If our official theories disagree with what we cannot help thinking outside the philosophy room, then no real equilibrium has been reached.

While common sense is not sacred (see Lewis 1986a, 134), it does constrain our theorizing. We should therefore expect a theory of chance to conform, broadly, to our core common sense beliefs about chance.

Another important aspect of Lewis's methodology is his treatment of theoretical terms. Lewis (1970) suggests that sentences of a theory employing theoretical terms (such as "chance has properties *F*, *G*, and *H*") can be replaced with existentially quantified sentences (such as "there exists an *x* with properties *F*, *G*, and *H*"). The denotation of the theoretical term (in this case, chance), is just whatever witnesses the existentially quantified sentence (or comes close to witnessing the sentence) – whatever plays the right role (or comes close).

Both of these points converge on a common idea: the chances are whatever plays the chance role. So to argue for the existence of deterministic chance, we should first identify the chance role, and then find something that can satisfy it while remaining compatible with determinism.

There is one complication, however. My chances are non-unique; they share the chance role with Lewis's chances. But Lewis (1970) claims theoretical roles cannot be shared: if a theoretical role has multiple realizers in the actual world, then the theory is defective. Fortunately, it's not clear that "chance" is exactly a theoretical term: there are quantum mechanical chances, statistical mechanical chances, economic and ecological chances. We might take "chance" to be a metatheoretical term, a role for terms in various theories. Or we might think of "chance" the way that Lewis (1980a, 126) thinks of "pain." There, Lewis claims that "pain" can refer either "non-rigidly" to the theoretical role or "rigidly" to the property that realizes it, and that "Nonrigidity might begin at home" – that is, there might be different realizers in different parts of the actual world. In the case of "pain," the relevant parts of the actual world are populations of creatures; in the case of "chance," we might think of them as different levels of description.

So the chance role can be shared. But what is the role? I will divide it into four parts: chance is indexed to time and a world, acts as a nomological magnitude and a Humean magnitude, guides credence, and arises in objectively undetermined situations.

18.2.1 Indexing

Lewis (1980b, 91) relativizes the chance function to both a time and a world. He illustrates the alleged time- and world-dependence of chance with the example of a traveler navigating a labyrinth of T junctions by flipping an objectively chancy coin every time she reaches a fork, and choosing her direction based on the coin's outcome. She begins at 11:00 am. As she moves through the labyrinth, a series of lucky and unlucky turns affect her chance of successfully reaching the center by noon. The chance of success rises and falls with her luck and, when she finally reaches the center at 11:59, the chance becomes 1, "then and forever more." In this example, the chances vary from time to time, and so seem to deserve a time index.

Just as the chances vary from time to time, they vary from world to world. If the coin had been more biased toward heads, the heroine's chance of reaching the center by noon would have been different. So the chances seem to deserve a world index.

Despite the appeal of the example, I will follow Meacham (2005) and Hoefer (2007) in rejecting Lewis's time- and world-dependent chances. This choice will require me to formulate the other theoretical roles of chance differently from Lewis and many of his followers, who frequently presuppose world- and time-dependence even when discussing independent matters.

I hold that chances should be indexed to background propositions. I will cash out this idea more fully in section 18.3; for now, some preliminary remarks will do. Chances are defined not *at* a time t in a world w, but given a background proposition B. Not just any proposition is a suitable background proposition. Rather, Ch_B is defined only if B is a proposition to the effect that a particular type of chance setup is instantiated at a particular space–time location.

We can think of Lewis's time- and world-indexing as a special case of proposition-indexing. Let H_{tw} be the complete history of w at t – "the proposition that holds at exactly those worlds that perfectly match w, in particular matter of fact, up to time t" Lewis (1980b, 97). And let L_w be the complete laws at w – the proposition that holds at exactly those worlds that have exactly the same laws as w. Then Lewis's background propositions are propositions of the form $H_{tw} \wedge L_w$: his chance setups are entire world histories together with the complete laws of nature.

Lewisians have at least good reasons to reject Lewis's time-indexed notion of chance. First, they ought to be skeptical about the existence of times, as distinct from places. Lewis believes in the space–time and relative simultaneity of relativistic physics, not the space and time of Newtonian physics (Lewis 1986a, 74; 1994, 474). As a good scientific realist, he should.

Second, the time-indexing of chance is closely linked to Lewis's idea that "what's past is no longer chancy" – that is, that no past event (as of t, in w) has a chance in the (0, 1) interval (as of t, in w). Suppose you accept my proposition-indexed paraphrase of Lewis's theory, which relativizes chances to propositions of the form H_{tw}. And suppose that the chance of any proposition, given itself, is 1. Then, given H_{tw}, the chance of anything entailed by H_{tw} is 1, while the chance of anything ruled out by H_{tw} is 0 (assuming these chances are defined at all). Since H_{tw} is a complete history, it either entails or rules out every proposition about the past (as of t). So no past event can have a chance other than 0 or 1 – what's past is no longer chancy.

But Lewisians should not let it come out as a conceptual truth that what's past is no longer chancy. Lewis (1973, 1979) held that it was metaphysically possible for the future to depend causally and counterfactually on the past. This is what happens in cases of time travel – for instance if I build a time machine to send a bucket of uranium back to the Triassic Period. But in cases of time travel,

shouldn't past events also take on chances in the (0, 1) interval? As I push the button on the time machine, it ought to be chancy whether all my uranium will have decayed by the end of the Jurassic Period. So Lewisians ought to reject the claim that chances are indexed to times.

Still, in introducing his time- and world-indices, Lewis has pinpointed a phenomenon in need of explanation. Chances are subject to *change* – they can differ from time to time. And chances are *contingent* – they can differ from world to world. Any proposition-relative account of chance must somehow account for these two phenomena.

18.2.2 Chance as a Nomological Magnitude

Lewis classifies chance as an implicitly modal concept, together with laws, causation, dispositions, and counterfactuals. The concept of chance is closely linked to some of these other concepts: in particular, to laws and causation.

First, chances figure in laws of nature. Lewis is particularly concerned with a class of laws he calls *history-to-chance conditionals*. A history-to-chance conditional is any proposition of the form $H_t \rightarrow \varphi(Ch_t)$, where H is a complete description of history up to time t, $\rightarrow$ is a strong conditional (Lewis suggests that a counterfactual conditional will serve), and $\varphi(Ch_t)$ is a proposition about the chances at time t.

Marc Lange (2006, 385) gives the following summary of Lewis's view of the relationship between chance and laws of nature.

> *Lawful Magnitude Principle (LMP)* If $Ch_{tw}(A) = x$, then the laws of w entail a history-to-chance conditional of the form: if the occurrent history of w through t is H, then $Ch_{tw}(A) = x$.

The LMP builds in assumptions about world- and time-indexing that I hope to reject. It does this in two ways: by explicitly indexing the chance function to a world–time pair, and by employing a concept of history-to-chance conditionals that links chances to particular times in particular worlds.

While a proposition-indexed variant of Lewis's theory ought to preserve some manner of close link between the chances and laws, it cannot preserve the letter of the LMP. But what, at bottom, is the LMP saying? I suggest that it says something like the following: It is the laws (perhaps with some supplementary contingent information) that determine the chances. The chances are what they are in virtue of the laws. The following principle, then, captures the spirit of the LMP without the time index.

> **LMP*** If $Ch_B(A) = x$ and B is true in w, then in virtue of the laws of w, $Ch_B(A) = x$.

Chance is linked with causation as well as laws. Lewis discusses this relationship in the 1986 appendix to Lewis (1973). Consider two events c and e, both of which, as a matter of fact, occur. Lewis holds that a sufficient condition for c's causing e is that if c had not occurred, then e's chance of occurring would have been significantly lower. (In such cases, we say that *e probabilistically causally depends* on c.)

Lewis takes pains to distinguish probabilistic causal dependence from deterministic causal dependence that obtains by chance. In a world where c and e both occur, *e deterministically causally depends* on c iff if c had not occurred, then e would not have occurred either. But Lewis holds that if e probabilistically causally depends on c, it is neither true that if c had not occurred, then e not have occurred, nor true that if c had occurred, then e would have occurred. Among the closest worlds where c occurs, there are some where e occurs, and some where e does not. So the counterfactual "if c had occurred,

then e would have occurred" is not true. Thus, if e probabilistically causally depends on c, e cannot exhibit deterministic causal dependence (chancy or otherwise) on c.

A satisfactory account of chance will secure the connection between chance and probabilistic causal dependence (not merely the kind of deterministic causal dependence that obtains by chance).

18.2.3 Chance as a Guide to Credence

Lewis famously held that chance constrains rational credence via the

> *Principal Principle (PP)*. Let C be any reasonable initial credence function. Let t be any time. Let x be any real number in the unit interval. Let X be the proposition that the chance, at time t, of A's holding equals x. Let E be any proposition compatible with X that is admissible at time t. Then
>
> $$C(A \mid X \wedge E) = x \text{ (Lewis 1980b, 87)}$$

Lewis's way of explicating the chance–credence link is idiosyncratic in a few ways. First, it presupposes that chances are indexed to times; we'll want to remove this presupposition.

Second, Lewis formulates the PP terms of reasonable initial credence functions. The idea is to abstract away from the contingent evidence that affects credence functions and consider a logical prior. 'Reasonable' is a permissive term: an initial credence function is reasonable iff it is rationally permissible. (Lewis explicitly allows for the possibility that many initial credence functions are reasonable.) An individual's credence function is reasonable iff it could be derived by conditionalizing a reasonable initial credence function on her total evidence.

Initial credence functions are an idealization – no real person has ever held one, and perhaps no person could. They are also mathematically unwieldy: since it would be mere prejudice to rule out genuine possibilities *a priori*, Lewis insists that every possibility should get non-zero initial credence. But there are uncountably many possibilities, so Lewis is forced to appeal in a hand-waving way to the mathematics of infinitesimals. It would be better to avoid the idealization if we could.

Luckily, we can. Let Cr be the reasonable credence function of an individual whose total evidence is the proposition E. Since reasonable credence functions come from reasonable initial credence functions by conditionalization on total evidence, there is some reasonable initial credence function C such that

$$Cr(A|Ch(A) = x) = C(A|Ch(A) = x \wedge E) = C(A|X \wedge E)$$

Recall that the PP requires that

$$C(A|X \wedge E) = x$$

where E is admissible. So what the PP requires of Cr is that

$$Cr(A|Ch(A) = x) = x$$

provided the individual's total evidence E is admissible.

The third idiosyncrasy in the PP is the concept of admissibility itself. Lewis (1980b, 92) glosses admissibility as follows: "Admissible propositions are the sort of information whose impact on credence about outcomes comes entirely by way of credence about the chances of those outcomes." Lewis holds that admissibility is implicitly relative to a time because the chance function is indexed to a time. Thau (1994) and Hall (1994) point out that admissibility should also be relativized to a

proposition in the domain of the chance function. Spelled out further, Lewis's gloss on admissibility says that propositions which are admissible at t with respect to A are the sort whose impact on credence about A comes entirely by way of credence about A's chance at t.

Leaving off the admissibility clause condemns the PP to either toothlessness or absurdity. The PP would be weak to the point of toothlessness if we claimed that for all reasonable initial credence functions C, propositions A, and real numbers x,

$$C(A|X) = x$$

This reformulation would leave us unable to derive information about the credence functions of most reasonable individuals with contingent information – our real target. A reasonable individual will often know a proposition E that is not entailed by information about the chances. Her credence Cr in A given X should equal x only if $C(A \mid X \wedge E) = x$, and without the admissibility clause, the PP does not speak to that.

On the other hand, the PP would be absurdly strong if it claimed that for all reasonable individual credence functions Cr, propositions A, and real numbers x,

$$Cr(A|X) = x$$

A will often be a knowable proposition, which means that some reasonable credence functions will belong to individuals who know A. By virtue of their reasonableness, these individuals will place credence 1 in A conditional on *anything*, including propositions that claim A's chance is between 0 and 1. Other reasonable credence functions will belong to individuals with excellent evidence for A – evidence that supports A irrespective of information about A's objective chance. These individuals will be justified in placing high conditional degrees of belief in A given propositions to the effect that A's chance is low. So it is crucial that the PP contain an admissibility clause in some form.

I propose the following version of the PP, revised to remove some of the idiosyncracies:

PP* Let Cr be any reasonable credence function held by someone whose total evidence is admissible with respect to A and Ch_B. Then

$$Cr(A|B) = Ch_B(A)$$

The PP* differs from the PP in several ways.

First, it appeals to reasonable individual credence functions instead of reasonable initial credence functions. Second, since PP* appeals to individual credence functions rather than initial credence functions, it puts the admissibility clause in a slightly different place – as a condition on what the agent already knows, rather than into the conditional probability as a condition the agent might conditionalize on. Third, it relativizes the chance function to a proposition instead of to a world and a time.

Finally, since I relativize chance to a proposition, PP* puts $Ch_B(A)$ on the right-hand side of the equation, rather than into the conditional probability on the left-hand side. I am assuming that if $Ch_B(A) = x$, then every reasonable individual knows *a priori* that $Ch_B(A) = x$; since chances are no longer world-relative, Ch_B does not depend on the answer to any empirical question. We could write the PP* as

$$Cr(A|B \wedge Ch_B(A) = x) = x$$

But what I've written is much the same, for on the right-hand side of the equation, x can be replaced with $Ch_B(A)$, while on the left-hand side, the condition $Ch_B(A) = x$ drops out.

What is the relationship between the correct chance–credence principle (be it PP or PP*) and the correct theory of chance? Lewis holds that in order to count as adequate, a theory of chance must support the link between chance and credence. But it's not entirely clear what the requisite support would amount to.

The earlier Lewis (1980b, 90) suggests that no analysis of chance is acceptable "unless it is *compatible* with the Principal Principle" (emphasis mine). This makes the requirement sound like one of logical consistency. But later, Lewis (1994, 484) writes, "Don't call any alleged feature of reality chance," he writes "unless you've already *shown* that you have something, knowledge of which could constrain rational credence" (again, emphasis mine). Here, the requirement sounds like one of justification – in addition to being logically consistent with the PP, an account of chance must provide us with some positive reason for adapting our credences to the chances.

What exactly is Lewis's requirement? He claims that the PP is a conceptual truth about chance,[1] which makes it puzzling that he should require any explanation at all. Strevens (1999) casts the problem as one of persuading a skeptic to adopt the PP, and argues that no theory of chance can satisfy it.

I will follow Hall (2004) in assuming that the problem is to derive the PP from other constraints on credence which we already accept. All other things being equal, it is better if a theory of chance can motivate PP in terms of its other commitments, rather than having to posit it as an additional constraint.

18.2.4 Chance as a Humean Magnitude

Lewis rejected the idea that chances were metaphysically basic. I will use the phrase *Humeanism about chance* for the view that the chances supervene entirely on non-chancy facts. Lewis was a Humean not just about chance, but about all nomological concepts, including laws, dispositions, counterfactuals, and causation. Since my focus here is chance, I will assume only that an adequate Lewisian theory is Humean about chance. The positive account I propose in 18.3 is neutral with respect to Humeanism about other nomological concepts.

Lewis defended a particular brand of Humeanism, the Best System account (BSA) of law and chance together. According to the BSA, laws of nature are theorems of the axiomatic system that does best at summarizing the distribution of non-nomological properties throughout the past, present, and future. "Best" is a matter of balancing theoretical virtues: simplicity, strength, and, in the case of indeterministic theories, *fit* – that is, assigning a higher chance to the actual world than to other possible worlds. (Remember that for Lewis the chances are lawful magnitudes in accord with the LMP.) But I take the BSA to be a substantive commitment of Lewis's, rather than a requirement on an adequate account of chance.

18.2.5 Objective Chanciness

"There is no chance without chance," Lewis (1980b) writes. "If our world is deterministic, there are no chances in it save zero and one." The underlying idea seems to be that an event could not count as chancy unless it really might have turned out differently. "Might have turned out differently" admits of either an epistemic or a metaphysical reading.

Bigelow, Collins and Pargetter (1993) and Schaffer (2003) propose formal principles that capture the metaphysical reading. Bigelow *et al.* (1993) suggest the

Basic Chance Principle (BCP) If $Ch_{tw}(A) > 0$, then there exists a world w_{ground} such that: (i) A is true at w_{ground}, (ii) w_{ground} matches w in occurrent history up to t, and (iii) $Ch_{twground}(A) = Ch_{tw}(A)$.

Schaffer (2003) endorses the BCP, but supplements it with the

Realization Principle (RP) If $Ch_{tw}(A) > 0$, then there exists a world w_{ground} such that: (i) A is true at w_{ground}, (ii) w_{ground} matches w in occurrent history up to t, and (iii) w_{ground} matches w in laws.

The BCP says that events with positive chances could have happened, given the actual chances and the proposition that generates them. The RP says that events with positive chances could have happened, given the laws and the proposition (in this case, a historical proposition) that generates the chances.

Both the BCP and the RP presuppose that the chances are indexed to times. However, we can furnish the following proposition-indexed principle, which replaces both of them at once.

RP* If $Ch_B(A) > 0$, then for every world w such that $\neg A \wedge B$ is true at w, there exists a world w_{ground} such that (i) $A \wedge B$ is true at w_{ground} and (ii) w_{ground} matches w in laws.

The RP* says that events with positive chances could have happened, given the background proposition (which specifies which chance setup is instantiated where) and the laws. In other words, if an outcome has a positive chance, given that a chance setup is instantiated, then it is nomologically possible for the chance setup to be instantiated and the outcome to occur.

It's important to note that Lewis himself rejects the BCP as a constraint on the concept of chance – his particular version of the BSA rules it out. While Lewis concedes that rejecting the BCP is "very peculiar" he maintains that it is "no worse than peculiar" (1980b, 483). He would doubtless say the same about RP. Nonetheless, both principles gesture at a metaphysical type of objective chanciness – one that would be worth having, if Lewisians could get it.

When Lewis himself discusses objective chanciness, he invokes epistemic notions. He suggests that the chances at a time t are closely associated with the *history–theory partition* at t. Two worlds are in the same cell of the history–theory partition for t iff they share both their history up to t and their complete theory of chance. The history–theory partition has three salient characteristics, one metaphysical and two epistemic.

1. It seems to be a natural partition, not gerrymandered. It is what we get by dividing possibilities as finely as possible in certain straightforward respects.
2. It is to some extent feasible to investigate (before the time in question) which cell of this partition is the true cell; but
3. It is unfeasible (before the time in question, and without peculiarities of time whereby we could get news of the future) to investigate the truth of propositions that divide the cells. (Lewis 1980b, 120–1)

Lewis concedes that the history–theory partition is not the only partition with these characteristics, but argues that deterministic "counterfeit" chances associated with other partitions are not chancy enough.

> Counterfeit chances will be relative to partitions; and relative, therefore, to standards of feasibility and naturalness; and therefore indeterminate unless the standards are somehow settled, or at least settled well enough that all remaining candidates for the partition will yield the same answers. Counterfeit chances

> are therefore not the sort of thing we would want to find in our fundamental physical theories, or even in our theories of radioactive decay and the like. (Lewis 1980b, 121)

There are two separate complaints in this passage. The first is that deterministic chances are not part of our fundamental physical theories. It's not clear that this complaint is true: Albert (2000), Loewer (2001, 2004), and Ismael (2009) argue persuasively that statistical mechanical probabilities are an irreducible and indispensable part of our fundamental physical theories. Even if the complaint is true, it's not clear why it should matter to whether there are deterministic chances: not all of objective reality is found in fundamental physics. Shoes, ships, and sealing wax are no less real for being non-fundamental.

The second complaint is more compelling. Defenders of deterministic chance need some positive account of how to arrive at the right set of chancy propositions. Lewis's demand for uniqueness might seem bizarre; he gives us not one partition, but a whole slew of them – one for every time. What business has he got demanding a unique partition from the defender of deterministic chance? Lewis's focus on epistemic considerations in this passage suggests that we need reassurance that, in all or most epistemic situations, there will be a unique chance function to guide credence. If there are multiple chance functions, we need a way of deciding which of them will serve as a guide. Lewis's time-indexing provides an easy answer: when two chance functions conflict, you should always defer to the one that's indexed to a later time. The defender of deterministic chance needs a different answer.

One more note: we needn't accept Lewis's assumption that the chances are associated with a partition. There may be propositions which have chances in some but not all possible worlds (a possibility that Lewis himself acknowledges). If so, then the set of propositions of the form $Ch_{\mathrm{B}}(A) = x$ or $Ch_{\mathrm{tw}}(A) = x$ will not be jointly exhaustive. The important aspect of the partition is that its members are mutually exclusive – that we never encounter conflicting chances.

18.3 A Theory of Deterministic Chance

I aim to explain how deterministic chances are possible on a broadly Lewisian theory. More specifically, I aim to explain how there can be deterministic chances that function as nomological magnitudes, guide credence, and arise in objectively chancy situations.

I do not claim that my deterministic chances are the only things capable of occupying the chance role. There is nothing in my account to rule out indeterministic chances, or deterministic chances of other sorts. I claim only that my deterministic chances are among the things that deserve to count as chances.

I begin with the concept of a *chance setup*: a repeatable type of event, such as a roulette spin or a coin toss. The chance setup has three ingredients:

1. A set of designated outcomes – all event types. In the case of the coin toss, these outcomes might be heads and tails. In the case of the roulette spin, they might be red and black, or outcomes consisting of the wheel's landing on particular numbers.
2. A range of input conditions – also event types. In the case of the coin toss, these might be combinations of initial upward velocity and initial angular velocities. In the case of the roulette spin, they might be initial angular velocities. I will assume that the input conditions come with a built-in measure, unique at least up to positive linear transformation. This assumption seems to hold good for velocity and angular velocity. I'll say more about where the measure might come from in section 18.3.3.

3. Chaotic dynamic laws mapping input conditions to outcomes. These laws will most likely not be fundamental laws, but rather *ceteris paribus* laws that characterize chancy systems stably situated in their environments. The initial speed of the roulette wheel determines the outcome only on the assumption that the roulette wheel is not suddenly stopped by the croupier, doused with mineral oil, destroyed in an explosion, or otherwise interfered with.

Each chance setup is associated with what I will call a *type-chance function*. Type-chance functions are indexed not to propositions, but to types of chance setups, and they assign probabilities not to propositions, but to types of outcomes. I will use lowercase letters for concepts associated with type-chance functions, so that ch_b is the type-chance function indexed to the experiment type *b*. Notice that the domain of the type-chance function will be limited. Although a given fair coin toss may result in outcomes of many types – a winning bet for someone who needs a new pair of shoes, Abraham Lincoln's nose pointing due north, the coin's scratching the table – the type-chance function associated with fair coin tosses will assign probabilities only to heads and tails.

For deterministic setups, I will define the type-chance of an outcome as the ratio between the measure of the input conditions that lead to that outcome and the measure of all the possible input conditions (according to the built-in natural way of measuring). For instance, suppose that we toss a fair coin, according to the setup in Figure 18.1. The chance of a heads outcome on this type of toss is roughly 1/2, because roughly 1/2 of the possible initial conditions lead to a heads outcome. (These are the conditions shaded black in the drawing.)[2]

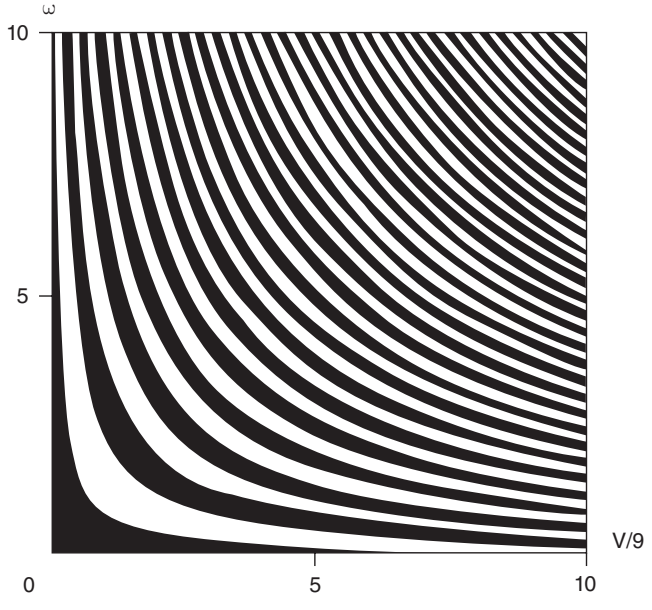

Figure 18.1 A diagram of a coin toss setup, from Diaconis (1998, 803). The *x* axis represents the coin's initial upward velocity v, while the y axis represents the angular velocity ω of its spinning. Assuming the coin begins with the heads side up, the black points represent initial conditions that lead to a heads outcome, while the white points represent initial conditions that lead to a tails outcome. Reprinted by kind permission of *Quarterly of Applied Mathematics*.

The next step is to get from type-chances of outcome types to token chances of propositions. Recall that the chances that ultimately interest us are indexed to propositions which specify that chance setups are instantiated at particular locations. A proposition B is associated with a chance function Ch_B iff B says that a chance setup of type b is instantiated at a particular spatiotemporal location l. Where a is an outcome type of chance setup b, and A is the proposition that a occurs at l, $Ch_B(A) = ch_b(a)$. For example, B might be the proposition that a fair coin toss occurs in a particular corner of the Coombs Building in Canberra on December 25, 2013. In that case, Ch_B will assign probability 1/2 to the proposition that a coin, tossed in said corner of the Coombs building on December 25, 2013, lands heads. This is because B specifies that the coin toss is a fair one (of a certain sort) and what it is to be a fair coin toss is just to have probability 1/2 of landing heads.

18.3.1 Indexing

My chances, unlike Lewis's chances, are neither explicitly time-indexed nor explicitly world-indexed. Nonetheless, my chances admit of both change and contingency.

They admit of change because chance setups are instantiated at locations in space and time. At different times, different chance setups will be salient, so different chance functions will be salient, so the chances will change. Lewis's example involves multiple token chance setups. All of these token chance setups have the same set of token outcomes: {the heroine reaches the center of the labyrinth by noon, the heroine does not reach the center of the labyrinth by noon}. And all of the token chance setups belong to similar types: they are navigations of a labyrinth (of a particular layout) by repeated tosses of a fair coin. But they do not belong to the same type, because each has a different range of input conditions – a different designated starting point in the labyrinth. And they are associated with different chances. As the character navigates the labyrinth, different chance setups become salient, and so the chances change.

My chances also admit of contingency, because they are indexed to contingent propositions. What makes a chance function Ch_B actual, rather than merely possible, is that B be true – that the relevant chance setup b be not instantiated at the relevant location l. Since B is contingent, Ch_B is contingent as well.

18.3.2 Chance as a Nomological Magnitude

My chances satisfy the LMP*. The laws governing the relationship between input conditions and outcomes make the chances what they are – they are built into the proposition B. Whatever the necessary conditions are for the chances to obtain in virtue of the laws, this is surely a sufficient condition.

My chances also undergird probabilistic causal dependence. Here it is necessary to rebut an argument by Lewis for the conclusion that probabilistic causal dependence requires indeterminism. In cases of deterministic "counterfeit" chance, he writes (in the postscripts to "Causation"),

> [the effect] e is after all predetermined either one way or the other, both in actuality and in the counterfactual situation without [the cause] c; but predetermined partly by details of prior historical fact that are far too minute to be discovered in advance. (Lewis 1986b, 183)

This, I claim, is too quick.

Lewis writes of *the* counterfactual situation without c, but as Edgington (2004, 14) points out, there need be no unique closest world where c does not occur. Instead, there may be many equally good ways for c not to occur.

> Our ordinary run-of-the-mill antecedents are not normally specific enough to be fed into deterministic laws. Even if coin-tossing is a deterministic process, no deterministic conclusion comes from the counterfactual supposition that you had tossed the coin, but only from a supposition of how *exactly down to the minutest detail* you tossed it.

One consequence of Edgington's observation is that deterministic worlds, like indeterministic worlds, can contain chancy causal dependence. Consider José, a denizen of a deterministic world who amasses a large fortune after wisely deciding not to squander his money on roulette. There is unlikely to be deterministic causal dependence between José's refusal to play roulette and his large fortune. Among nearby possible worlds where José does play roulette, there will be some where he bets in just such a way, and the croupiers spin their wheels at just such a speed, to ensure that he wins back as much as he loses. José's wealth causally depends on his refusal to play roulette, but only probabilistically. It's not true that had he played roulette, he would have lost money, but nonetheless, it's true that had he played roulette, he would *probably* have lost money.

Like causation and counterfactual dependence, chances lack a built-in temporal direction. But like causation and counterfactual dependence, chances typically have a temporal direction. Does this phenomenon have an explanation?

Alas, explaining the temporal direction of any nomological concept is difficult and tricky business. But we can at least hope that the explanation of the temporal direction of chance is linked to the explanation of the temporal direction of causal and counterfactual dependence. Here is one link. To some extent, one can directly intervene on the input conditions of a chance setup while leaving the chance setup intact. (For instance, one can yell at the croupier to spin harder, or jostle the hand of the coin-tosser, without affecting either the connection between input conditions and output conditions or the range of possible input conditions.) But one cannot intervene with the outcome of a chance setup while leaving its links with the input conditions of the chance setup intact. This suggests that there is a causal distinction between input conditions and outcomes that mirrors the direction of causation in general.

18.3.3 Chance as a Guide to Credence

Strevens (2003) suggests a method that allows us to derive the PP from other constraints. Take any probability density function f over the input conditions of a chance setup b that is *locally close to uniform* – that is, one that varies only slightly across small changes in input conditions. Let i be a variable ranging over input conditions.[3] (In the roulette wheel case, we would have a variable measuring the input speed of the wheel.) Let $c_a(i)$ be the characteristic function of input conditions that get mapped to outcome a by the dynamic laws. (In the roulette wheel case $c_{red}(i)$ is the function that equals 1 if i is an angular velocity that will result in a red outcome, and 0 otherwise. Strevens argues that insofar as the dynamics of the chance function are chaotic in the right way, $ch_b(a) \approx \int f(i)c(i)di$ – in other words, that the chance of an outcome is roughly equal to the outcome's probability (according to the density function).

Furthermore, the situation generalizes to situations with chancy dynamic laws. Let $p_a(i)$ be the probability, according to the laws, that an instance of the setup in initial condition i will yield outcome a. If the dynamics of the chance setup are chaotic in the right way, they will furnish a number $ch_b \approx \int f(i)c(i)di$. Once again, Strevens argues that $x = ch_b(a)$ – that the chance of an outcome is roughly equal to its probability (according to the density function).

Suppose we interpret the density functions f subjectively, as giving credences over input conditions. (This is the suggestion made by Savage 1973.) We might constrain reasonable credence as follows: an agent who knows that a chance setup b is instantiated, and has no inadmissible information,

should have locally close-to-uniform credences about the input conditions of this instance of *b*. Then we can put the following interpretation on Strevens's result: For an agent who knows that *B* and has no inadmissible information,

$$Cr(A) \approx Ch_{\mathrm{B}}(A)$$

Since agents should conditionalize on their total evidence, an agent who doesn't know that *B*, and has no inadmissible information relative to *B* should set

$$Cr(A|B) \approx Ch_{\mathrm{B}}(A)$$

In other words, agents with suitable credence functions about the input conditions of chance distributions must satisfy the PP (or at least come close).

This idea also suggests a way of cashing out the concept of admissible information. Admissible information (relative to a proposition *B* about the instantiation of a chance setup) is information that does not tell us too much about which of *B*'s input conditions is actual – information that would not lead a reasonable person's credences to deviate too far from local uniformity over input conditions. "Too much" and "too far" are vague, so perhaps a better way to put the point is that information is inadmissible to the extent that it justifies non-uniform credences over input conditions. Since the outcome proposition A is already built in to the proposition *B* about the instantiation of a chance setup, there is no need to relativize admissibility to it.

Where does the uniformity constraint on credence about input conditions come from? I see three possible answers: the credence functions may come from objective symmetries, from smoothed-out frequencies, or from a universal ur-chance distribution of the sort defended by Barry Loewer. Rather than attempting to adjudicate among these answers, I will lay them out and gesture at their strengths and weaknesses.

Symmetries Perhaps the measure over input conditions is somehow metaphysically special – in Lewis's terms, perhaps it is uniquely *natural*. If so, then a reasonable principle of indifference might require us to assign equal credences to equal-sized chunks of state space.

Lewis would not be happy with the symmetry-based answer; he holds that "symmetries are only defeasible constrainers of rational credence" (Lewis 1994, 476). But, based on the evidence Lewis cites, it's not obvious that this is the right conclusion. He goes on to consider the example of a drunkard navigating a maze of symmetrical T-junctions. Although the left–right symmetry of the T-junction might initially justify in our placing equal credence in the proposition that a given drunkard is about to turn left and the proposition that he is about to turn right, "The symmetry of the T-junctions would no longer require 50–50 division of credence if we knew that, despite this symmetry, the drunkards also turn right nine times out of ten."

But the example admits of more than one explanation. Some symmetries are superficial, and not borne out by any underlying natural measure. If drunkards come equipped with some hidden inner mechanism that causes them to favor right-hand turns, this would break the apparent symmetry between left and right turns. It is not always easy to tell whether a symmetry is spurious or deep. A defender of the symmetry-based answer might claim that frequencies at odds with the symmetry are a source of evidence that the symmetry is spurious rather than deep.

A well-known problem with symmetry-based accounts of chance is the cluster of objections known as *Bertrand's Paradox*, all of which turn on the difficulty of picking out natural measures for quantities. The success of the symmetry-based strategy will depend on its ability to rebut Bertrand's Paradox by uncovering natural measures.

Frequencies Those who reject the symmetry-based account might claim that the measure over input conditions is a slightly altered frequency measure. This is a view that Strevens (2003) proposes, but does not explicitly endorse. The qualifier "slightly altered" is important, because in a finite world, we do not want all the probability concentrated at a finite number of points in a continuous range of input conditions. Strevens suggests a "smoothing-out" procedure that spreads the probability of each actual value of the input parameters over a range of nearby values.

Ur-Chance A third possibility is to adopt a proposal by Barry Loewer (2001, 2004). Loewer endorses the BSA, and argues that the best system should include not just dynamic laws mapping earlier states of the universe deterministically or stochastically onto later ones, but also a probability distribution over initial states of the universe. This initial distribution furnishes chances conditional on any contingent proposition B: all you have to do is conditionalize the initial distribution on the contingent proposition on B.[4] So in particular, where B says that a chance setup is instantiated at a particular location, Loewer's account automatically furnishes a chance conditional on B.

Why is this chance a guide to credence? Loewer (2004, 1122) suggests that the guidance is built into the relevant concept of *best* in "best system."

> Fit can be understood as a kind of informativeness – the information that probabilistic propositions provide concerning the propositions they attribute probability to . . . But these probabilities are informative only to someone who is willing to let them constrain her degrees of belief.

If the BSA turns out to be right, then Loewer's account is the right one for Humeans to adopt. A complete defense of the BSA would include a way of spelling out the theoretical virtues in a way that makes them accessible to human minds, yet mind-independent, and it is an open question whether this can be done. Even if the BSA fails, or the best system turns out not to include the right kind of probability distribution over initial conditions of the entire world, Lewisians can avail themselves of the symmetry- or frequency-based proposal.

18.3.4 Chance as a Humean Magnitude

My account is Humean about chance, in the weak sense I proposed in section 18.2.4: the chances supervene on the non-chancy facts. However, my chance setups are made partly from modal ingredients: laws together with ranges of possible input conditions. Lewisians can easily combine my account with a BSA of laws, or of laws and microphysical chances together. But what should they make of the "possible" in "possible input conditions"?

There are two potential readings. One is epistemic: the possible input conditions to a given chance setup, for a given agent, are simply the input conditions that for all she knows obtain. The other reading is agentive: the possible input conditions to a given chance setup are the ones that its causal ancestors can produce. The input conditions to the roulette wheel are all the speeds at which the croupier can spin it, the input conditions to the coin are the various degrees of force with which the tossing device can toss it, and so forth. If we are lucky, we will not have to choose: the agentive and epistemic possibilities will line up.

18.3.5 Objective Chanciness

My deterministic chances are objectively chancy both in the metaphysical sense and in the epistemic sense I outlined above. Recall that I cashed out the metaphysical sense of objective chanciness in terms of the

RP* If $Ch_B(A) > 0$, then for every world w such that $\neg A \wedge B$ is true at w, there exists a world w_{ground} such that (i) $A \wedge B$ is true at w_{ground} and (ii) w_{ground} matches w in laws.

It is easy to find a world w_{ground} to witness the RP*: any world that combines the chance setup b with an appropriate input condition will do.

For objective epistemic chanciness, we want setups whose chances are at least somewhat feasible to ascertain, but whose outcomes are not. Furthermore, we want these setups to be individuated in a somewhat natural way. Coin tosses, roulette wheels, and thermodynamic systems are all paradigm cases of such setups. Furthermore, our inability to discern the exact speed of a roulette wheel or the exact microstate of a thermodynamic system is no accident. Roulette wheels spun at similar speeds, and systems with the same thermodynamic properties and similar microstates, genuinely resemble each other.

So propositions characterizing deterministic chance setups behave like the members of Lewis's history–theory partition: they are natural, easy to discriminate amongst, and difficult to divide up into finer-grained propositions in a way that's relevant to credences about outcomes. There is only one question: are we likely to face overlapping chance setups in situations where there is no way of breaking ties?

Unlike Lewis, we have no logical or metaphysical guarantee against this possibility. But we can hope that nature is kind to us. In most plausible actual situations where a chance setup belongs to multiple types, one type clearly takes precedence over the others. The same trial may be both *a fair coin toss* and *a fair coin toss with a particular initial angular velocity and a particular upward velocity*. Someone who knows that the coin toss belongs to both types should use the second type to guide her credences. There is an important sense in which the second type of trial is a narrowing of the first: both types of trials are associated with the same dynamic laws and outcomes, but the range of input conditions for the second type is a proper part of the range of input conditions for the first. When a type of chance setup b' is a narrowing of another type b in this sense, the chances associated with b' will trump the chances associated with b. If there are trials that belong to multiple types, none of which is a narrowing of the other, then this spells trouble for deterministic chance. But we can hope that nature is kind.

18.3.6 Conclusion

At the end of "Humean Supervenience Debugged," Lewis (1994) writes, "A feature of Reality deserves the name of chance to the extent that it occupies the definitive role of chance…But near enough is good enough." Lewis is addressing a different issue from the one at play here – he is responding to an argument for the conclusion that the BSA is incompatible with the Principal Principle. But the basic idea still stands. Anything that comes close to satisfying Lewis's platitudes about chance, and does so by satisfying his broader theoretical requirements, ought to count as a chance.

It is true that my deterministic chances, unlike Lewis's chances, are not time-indexed. It is true that they do not exactly satisfy principles, proposed by Lewis and others in a broadly Lewisian tradition, that presuppose time-indexing. Nonetheless, they satisfy appropriately proposition-indexed versions of various platitudes about chance. By Lewis's own lights, deterministic chances behave like chances: they change, they are closely connected to nomological concepts like causation and laws, they guide credence in accord with the PP, they supervene on non-chancy matters of fact, and they emerge in objectively chancy situations. Lewisians should love them.

Notes

1 "For I think the only basic principle we have about chance, the principle that tells us all we know, is the Principal Principle" (Lewis 1994, 485). Lewis is no doubt exaggerating when he claims the PP is all that we know about chance (see Hall 2004; Schaffer 2003), but it is clear that he considers it to be central to the role of chance.

2 For indeterministic setups, the definition of type-chance is slightly more complex, but in the same spirit. I assume the laws assign, for each input condition, a nomological probability distribution over possible outcome conditions. Suppose we have a function that maps each input condition to the nomological probability of the outcome on the input condition. (The indeterministic version of Figure 18.1 would have bands of various shades of gray rather than black and white, where darker bands indicate a higher physical probability of heads.) Integrating this function with respect to the input variable yields the chance of the outcome. (Even though the black ink is distributed more diffusely, the chance of heads is still the proportion of the area that can be colored black with all the black ink to the total black area.)

3 I assume here that the input conditions can be measured using a single variable; Strevens explains how to generalize the account to multivariable cases.

4 Well, not strictly any proposition – the non-measurable ones are still out.

References

Albert. D.Z. (2000). *Time and chance*. Cambridge, MA: Harvard University Press.

Bigelow, J., Collins, J. & Pargetter, R. (1993). The big bad bug: what are the Humean's chances? *British Journal for the Philosophy of Science*, 44(3), 443–62.

Diaconis, P. (1998). A place for philosophy? The rise of modeling in statistical science. *Quarterly of Applied Mathematics*, 56(4), 797–805.

Edgington, D. (2004). Counterfactuals and the benefit of hindsight. In P. Dowe & P. Noordhof (Eds.), *Cause and chance: causation in an indeterministic world*. London: Routledge.

Hall, N. (1994). Correcting the guide to objective chance. *Mind*, 103(412), 505–18.

Hall, N. 2004). Two mistakes about credence and chance. *Australasian Journal of Philosophy*, 82(1), 93–111.

Hoefer, C. (2007). The third way on objective probability: a skeptic's guide to chance. *Mind*, 116, 549–95.

Ismael, J. (2009). Probability in deterministic physics. *Journal of Philosophy*, 106 (2), 89–108.

Lange, M. (2006). Do chances receive equal treatment under the laws? or: Must chances be probabilities? *British Journal for the Philosophy of Science*, 57(2), 383–403.

Lewis, D. (1970). How to define theoretical terms. *Journal of Philosophy*, 67(13), 427–46.

Lewis, D. (1973). Causation. *Journal of Philosophy*, 70(17), 556–67. Reprinted with postscript in Lewis (1986b), pp. 160–213.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, 13(4), 455–76. Reprinted with postscript in Lewis (1986b), pp. 32–52.

Lewis, D. (1980a). Mad pain and martian pain. In N. Block (Ed.), *Readings in the philosophy of psychology*, Vol. I, pp. 216–22. Cambridge, MA: Harvard University Press. Reprinted with postscripts in Lewis (1983), pp. 122–32.

Lewis, D. (1980b). A subjectivist's guide to objective chance. In R. Jeffrey (Ed.), *Studies in inductive logic and probability*, Vol. II. Oakland, CA: University of California Press. Reprinted in Lewis (1986b), pp. 83–113.

Lewis, D. (1983). *Philosophical papers*, Vol. I. Oxford: Oxford University Press.

Lewis, D. (1986a). *On the plurality of worlds*. Oxford: Blackwell .

Lewis, D. (1986b). *Philosophical papers*, Vol. II. Oxford: Oxford University Press.

Lewis, D. (1994). Humean supervenience debugged. *Mind*, 103, 473–90.

Loewer, B. (2001). Determinism and chance. *Studies in History and Philosophy of Science Part B*, 32(4), 609–20.

Loewer, B. (2004). David Lewis's Humean theory of objective chance. *Philosophy of Science*, 71(5), 1115–25.

Meacham, C. (2005). Three proposals regarding a theory of chance. *Philosophical Perspectives*, 19, 281–307.
Savage, L.J. (1973). Probability in science: a personalistic account. In P. Suppes (Ed.), *Logic, methodology and philosophy of science IV*, pp. 417–28. Amsterdam: North Holland.
Schaffer, J. (2003). Principled chances. *British Journal for the Philosophy of Science*, 54(1), 27–41.
Schaffer, J. (2007). Deterministic chance. *British Journal for the Philosophy of Science*, 58, 113–40.
Strevens, M. (1999). Objective probability as a guide to the world. *Philosophical Studies*, 95(3), 243–75.
Strevens, M. (2003). *Bigger than chaos: understanding complexity through probability*. Cambridge, MA: Harvard University Press.
Thau, M. (1994). Undermining and admissibility. *Mind*, 103(412), 491–504.

19

Lewis on Causation

CHRISTOPHER HITCHCOCK

19.1 Introduction

David Lewis produced several different versions of a counterfactual theory of causation over the course of his career, beginning with the second paper in his first undergraduate philosophy class (as reported in Lewis 2004a). The proposal reached a wider audience in 1973, when his paper "Causation" appeared in the *Journal of Philosophy* (Lewis 1973a). The paper was published in advance of being presented at the 1973 meeting of the American Philosophical Association (Eastern Division). The published paper was shortened from the longer version that had been submitted to the APA program committee. This paper was reprinted in Volume II of Lewis's *Philosophical Papers* in 1986 (Lewis 1986a), along with extensive postscripts (Lewis 1986b). Some of the material in the postscripts was drawn from the longer version of the original paper; additional material was taken from an epilogue to a German translation of "Causation" appearing in 1981 (Lewis 1981a). The postscripts extended and revised the original theory in a number of ways. In 1999, he offered a very different theory in his Whitehead lectures at Harvard. An abridged version was published as "Causation as Influence" in the *Journal of Philosophy* (Lewis 2000); a more extended version appeared posthumously in 2004 (Lewis 2004a).

The details of any of Lewis's proposed analyses are only a small part of the lasting legacy of his contributions to the understanding of causation. To a large extent, that legacy is in the conceptual framework for exploring the connections between causation and counterfactuals, in the conceptual tools that were developed to defend the analyses and fashion new ones, in the problems and questions raised by critics, and in the identification and classification of different kinds of causal structure. In this broad sense, the impact of Lewis's work on causation is colossal, extending well beyond philosophy and into diverse fields such as legal theory, psychology, statistics, epidemiology, and artificial intelligence.

Causation played a prominent role in Lewis's program of Humean supervenience, first articulated in Lewis (1980) and Lewis (1986a, Introduction). He thought that causation could be reduced, via a series of intermediary concepts,[1] to the "Humean mosaic" – a "vast mosaic of local matters

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

of particular fact" (Lewis 1986a, ix) – of the actual world and of other possible worlds. Causation, in turn, could be used in analyses of a variety of other concepts: persistence through time, survival, personal identity, dispositions, mental content, perception, linguistic meanings, rational decision, and explanation. Not all of these concepts are part of metaphysics proper, so perhaps they do not fit into the program of Humean supervenience. Nonetheless, causation for Lewis is clearly a central nexus in an elaborate network of concepts. I have attempted to reconstruct a part of this network in Figure 19.1. Many of the arrows in this diagram will be explained in other chapters of this volume. The focus of this chapter will be on the connection between counterfactuals and causation, and on the use of causation in the analyses of other concepts, especially decision and dispositions.

Lewis's methodology in analyzing causation was primarily the traditional "Platonic" model of conceptual analysis. Proposed definitions of causation are tested by comparing their verdicts with those of common sense or intuition in a variety of hypothetical cases. If we are told that Billy and Suzy threw stones at a bottle, that Suzy's stone arrived at the bottle before Billy's could reach it,

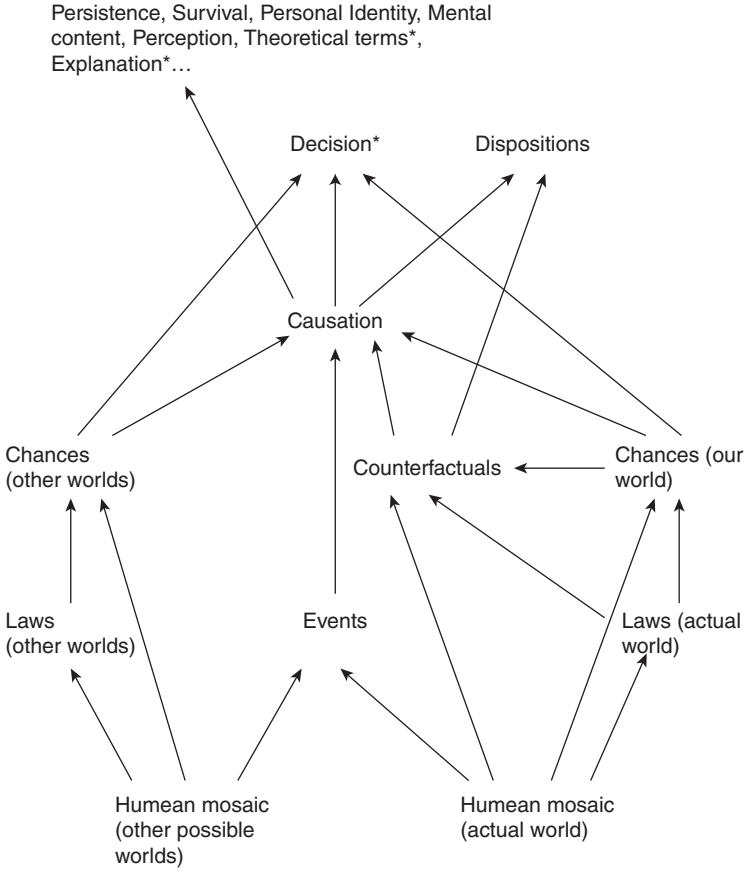

Figure 19.1 The place of causation in Lewis's program of Humean supervenience. Concepts that are not part of metaphysics proper are marked with an asterisk. Created by Christopher Hitchcock.

and that the bottle shattered shortly after impact, common sense says that Suzy's throw caused the bottle to shatter, while Billy's did not.[2] When a definition of causation is applied to this case, it had better reach the same verdict. Lewis wrote:

> When common sense delivers a firm and uncontroversial answer about a not-too-far-fetched case, theory had better agree. If an analysis of causation does not deliver the common-sense answer, that is bad trouble. (1986b, 194)

> When our opinions are clear, it's incumbent on an analysis of causation to get them right. (2000, 182; 2004a, 80)

But Lewis didn't take this approach to every topic. Elsewhere, he was willing to embrace counterintuitive consequences to meet theoretical needs. The most famous example is no doubt his embrace of modal realism – a belief in the reality of alternative possible worlds. Closer to the present topic, Lewis explicitly fashioned a theory of events (Lewis 1986c) and his criteria for similarity among possible worlds (Lewis 1979) to meet the needs of his theory of causation, rather than to capture common sense judgments on these matters. The tension between these two desiderata of analysis will be a central theme in my discussion. Specifically, I will argue that the concept Lewis attempted to analyze was ill suited to the project of providing causal analyses of further concepts.[3] Indeed, rather surprisingly, Lewis rarely appealed to his own definitions of causation when attempting such analyses. This suggests a reorientation: an approach to causation that is still very much Lewisian insofar as it places causation in the same position in the nexus of concepts, but focuses more heavily on the interplay between definitions of causation and the further theoretical needs such a concept will be expected to meet. Such an approach gives rise to new questions and new heuristics, and provides a framework for carrying Lewis's framework forward.

19.2 Preliminaries

Before turning to Lewis's analyses of causation, I will briefly review two preliminary pieces of conceptual apparatus. First, we need an account of events. Lewis (1986c) is his most detailed treatment. An event is a class of spatiotemporal regions of possible worlds. If an event occurs in the actual world, it will include a spatiotemporal region within the actual world. This region specifies when and where the event occurs, how long it lasts, and what its spatial boundaries are. Spatiotemporal regions in other possible worlds specify what would count as the occurrence of the very same event. An event may be more or less *fragile*. An event is fragile to the extent that it would not have occurred if things had been different in various ways. A fragile event must occur in some very specific way, or else not occur at all. Suppose that Suzy and Billy share a passionate kiss. We could have two different events. One, which we will call "the kiss," would have occurred even in a G-rated world where Suzy gives Billy a chaste peck on the cheek. A second, "the passionate kiss," would not have occurred in such a world. In this respect, the passionate kiss is more fragile that the kiss.

In order to be related as cause and effect, two events must be *distinct*. Distinctness is stronger than mere non-identity. Two events fail to be distinct if they are logically or mereologically related. In our example, the kiss and the passionate kiss are not distinct: the kiss will occur in any possible world in which the passionate kiss occurs. The passionate kiss will also fail to be distinct from "the courtship," a more temporally extended event that includes the passionate kiss and many others like it.

Second, while counterfactuals are notoriously context-sensitive, Lewis held that there was a standard resolution of counterfactuals involving the occurrence or non-occurrence of particular events,

according to which they do not "backtrack." Suppose that as a result of their kiss, Billy's cheeks flushed, and Suzy's pulse quickened. Then the following counterfactuals would be true:

If Suzy had not kissed Billy, his cheeks would not have flushed.
If Suzy had not kissed Billy, her pulse would not have quickened.

But the following counterfactuals are false:

If Billy's cheeks had not flushed, Suzy would not have kissed him.
If Billy's cheeks had not flushed, Suzy's pulse would not have quickened.

When we introduce a hypothetical change, such as supposing that the kiss had not occurred, the effects of this change propagate forward. But effects of a hypothetical change, such as the absence of excess blood from Billy's cheeks, do not propagate backward to the kiss, nor do they propagate backward and then forward again (hence "backtracking") to other effects of this kiss.

According to Lewis (1973b), a counterfactual is true if its consequent is true in all of the "closest" possible worlds where the antecedent is true. The closeness of two worlds depends upon their overall similarity. Lewis (1979) develops an account of the criteria for similarity that is intended to secure the needed asymmetry in the truth values of counterfactuals.

Finally, it will be helpful to have some notational shorthand. We will write '$A \rightarrow C$' for the counterfactual 'If A were true, then B would be true'. If e is an event, we will write '$O(e)$' for the proposition that e occurs.

19.3 Lewis's Analyses

I will divide Lewis's treatment of causation into three stages, labeled by the dates of Lewis's most significant publications on causation. The first will be the theory presented in the 1973 paper "Causation." The second will include the amendments included in the postscripts to "Causation" in *Philosophical Papers*, Volume II, in 1986. I will label these with the date 1986, even though some of them actually date back to the time of the original paper. Finally, I will present the theory from "Causation as Influence," first published in 2000.

19.3.1 Causation 1973

Lewis (1973a) arrives at an analysis of causation through a chain of definitions (page numbers refer to the reprinted version in 1986a):

> **Counterfactual Dependence:** Let $A_1, A_2, \ldots$ be a family of possible propositions, no two of which are compossible; let $C_1, C_2, \ldots$ be another such family (of equal size). Then if all the counterfactuals $A_1 \rightarrow C_1, A_2 \rightarrow C_2, \ldots$ between corresponding propositions in the two families are true, we shall say that the C's *depend counterfactually on the* A's. (164–5)

> **Causal Dependence (families of events):** Let $c_1, c_2, \ldots$ and $e_1, e_2, \ldots$ be distinct possible events so that no two of the c's and no two of the e's are compossible . . . [T]he family $e_1, e_2, \ldots$ of events

> *depends causally* on the family $c_1, c_2, \ldots$ iff the family $O(e_1), O(e_2), \ldots$ of propositions depends counterfactually on the family $O(c_1), O(c_2), \ldots$ (166)

The families of events $c_1, c_2, \ldots$ and $e_1, e_2, \ldots$ are like two distinct variables. For example, the events $c_1, c_2, \ldots$ might involve Billy depressing the accelerator of his car by different amounts, while $e_1, e_2, \ldots$ are events in which Billy's car travels at different speeds. The counterfactuals $c_1 \rightarrow e_1$, $c_2 \rightarrow e_2$, capture the way in which the speed of the car depends upon the depression of the accelerator.

I think that it would be useful to loosen this definition of causal dependence among families of events in two ways. First, the requirement that *all e*'s be distinct from one another is too strict. Suppose that Suzy, a trained opera singer, is singing outside Billy's window. As she sustains a note, his window vibrates. It may be that the amplitude of the vibration depends upon the frequency, or pitch, of the note. Moreover, it may be that the vibration would reach a maximum amplitude at some specific pitch (perhaps F-sharp below high C), falling off for higher and lower pitches. In this case, the pitches don't map one-to-one onto amplitudes, but it still seems reasonable to say that the amplitude causally depends upon the pitch. Second, the families of events should be allowed to include the absence of an event of the relevant kind – for example, Suzy's not singing at all, Billy's window not vibrating at all. Such absences are not themselves events, but the corresponding propositions are unproblematic. Thus the families of propositions that stand in a relationship of counterfactual dependence can include not only propositions of the form $O(c_1), O(c_2), \ldots$ and $O(e_1), O(e_2), \ldots$, but also propositions of the form 'none of the events $c_1, c_2, \ldots$ occurs' and 'none of the events $e_1, e_2, \ldots$ occurs.' This looser definition of the causal dependence among families of events more closely fits with his later applications of the concept.

As we shall see, I think that this notion of causal dependence between families of events is really the fundamental causal notion, rather than causation itself. That is why I am taking this somewhat circuitous route to arrive at Lewis's 1973 definition of causation.

The chain of definitions continues:

> **Causal Dependence (single events):** Let c and e be two distinct possible particular events. Then e *depends causally* on c iff the family $O(e), \sim O(e)$ depends counterfactually on the family $O(c), \sim O(c) \ldots$ The dependence consists in the truth of two counterfactuals: $O(c) \rightarrow O(e)$ and $\sim O(c) \rightarrow \sim O(e)$. (1986a, 166–7)

If c and e both occur, the first counterfactual is automatically true (by the centering property of counterfactuals; Lewis 1973b). Causal dependence among single events is a special case of causal dependence among families of events if the latter concept is weakened in the second of the ways I recommended above. It is not a special case of causal dependence among families of events as Lewis originally defined it, since the propositions $\sim O(c)$ and $\sim O(e)$ are not propositions saying that some event occurred.

If events c and e both occur, causal dependence of e on c is sufficient, but not necessary, for c to be a cause of e. Lewis defines causation (finally!) as the *ancestral* or *transitive closure* of causal dependence between events that occur. Consider the relation x is a parent of y. This relation is not transitive. If x is a parent of y, and y is a parent of z, it does not follow (and is typically not the case) that x is a parent of z. The relation x is an ancestor of y is the transitive closure of the parenthood relation (hence the term "ancestral"). x is an ancestor of y just in case x is a parent of y, or x is a parent of a parent of y, or x is a parent of a parent of a parent of y, or ... Thus we have:

> **Causation (1973):** Event c is a cause of event e iff there is a chain of events, $c = d_1, d_2, \ldots, d_n = e$, all of which occur, such that for each $i = 1, \ldots, n - 1$, d_{i+1} causally depends upon d_i.[4]

Substituting for defined terms gives us:

> **Causation (1973):** Event c is a cause of event e iff there is a chain of events, $c = d_1, d_2, \ldots, d_n = e$, such that (i) each of these events occurs; (ii) all of these events are distinct from one another; (iii) for each $i = 1, \ldots, n - 1$, the following counterfactual is true: if d_i had not occurred, d_{i+1} would not have occurred.

It is common in presentations of Lewis's 1973 theory to bypass the earlier definitions and go straight to this final definition.

The final step of taking the transitive closure of causal dependence has two consequences. The first and most obvious is that it makes causation transitive: If c is a cause of d, and d is a cause of e, then c is a cause of e. While a number of authors have argued that causation is not transitive (see, e.g., McDermott 1995 for several purported counterexamples), Lewis always maintained the transitivity of causation. Lewis (2004a) contains a sustained defense of transitivity. All of Lewis's definitions of causation involved a similar final step of taking the transitive closure of some relation.

A second consequence is that it helps the account deal with certain kinds of causal *preemption*. Suzy throws her rock at a bottle; the rock hits the bottle squarely, and the bottle shatters. Billy is standing ready with his rock. As soon as he sees Suzy throw, Billy drops his rock (knowing that Suzy has perfect aim). But if Suzy hadn't thrown her rock, Billy would have thrown his, with the same destructive effect. Billy and his rock constitute a backup causal process that is *preempted* by Suzy's throw. In this example, Suzy's throw causes the bottle to break, but the breaking does not causally depend upon the throw; hence causal dependence is not necessary for causation. However, consider an intermediate event: Suzy's rock in flight toward the bottle. If Suzy hadn't thrown, the rock would not have been in flight. Thus the rock's flight causally depends upon Suzy's throw. Now suppose that the rock had not been in flight. Since counterfactuals do not backtrack, Suzy still would have thrown (the rock miraculously vanishing, falling to the ground, or some such), and Billy still would have dropped his rock. Hence, if the rock had not been in flight, the bottle would not have shattered. Here there is a chain of causal dependence running from Suzy's throw, to the rock in flight, to the shattering of the bottle.

19.3.2 Causation 1986

The postscripts to "Causation" published in 1986 (Lewis 1986b) contained three different modifications to the theory. First, he extended the definition of causal dependence between single events to allow for indeterministic causation. This involved weakening the definition of causal dependence among single events.

> **Causal dependence (chancy; single events):** Suppose that events c and e are distinct, and both occur. Moreover, suppose that c occurred at time t, and that the objective chance at t that e would occur was x, that is, $ch_t(e) = x$. Then e causally depends upon c iff $\sim O(c) \rightarrow ch_t(e) \leq y$, where $y \ll x$.

That is, if c had not occurred, the chance of e would have been considerably lower than it actually was. The inequality in the consequent permits the chance of e to be different in nearby possible worlds where c does not occur, so long as the chance is substantially less than x in all of them. The earlier definition of causal dependence among single events is the special case where $y = 0$. Just as in the original theory, causation is the ancestral of causal dependence. This allows the theory to accommodate cases of preemption similar to that described in the previous section.

Lewis later abandoned this account of indeterministic causation (in Lewis 2004a). Suppose that Suzy (now an imperfect shot) throws a rock at the bottle, thus increasing the chance that it will break. Her rock misses, but the bottle shatters anyway (perhaps due to Billy's throw). Lewis's account incorrectly rules that Suzy's throw is a cause of the shattering.[5]

The second modification to appear in the 1986 postscripts was an extension to cases of prevention, causation by omission, and other kinds of case where the cause or effect is not an event, but the absence of an event. When Billy and Suzy go for a drive in the hills, it may be that Suzy's taking Dramamine prevents her from getting carsick, that is, causes her not to get carsick. Similarly, Billy's failure to take Dramamine may be a cause of his carsickness. Suzy's not getting carsick and Billy's not taking Dramamine are not events; indeed, according to Lewis (1986b), they are not anything at all. It follows that causation is not really a relation: sometimes the relata go missing. Nonetheless, the relevant counterfactuals seem straightforward: if Suzy had not taken Dramamine, she would have been carsick; if Billy had taken Dramamine, he would not have been carsick. There remain some issues about precisely how to formulate the expanded theory, but Lewis left these details as an open problem. Lewis continued to maintain that absences of events could be causes and effects when he reformulated his theory once again in the late 1990s (Lewis 2000, 2004a, 2004b).

The third, and most dramatic, modification was motivated by a new kind of preemption case. Suppose that Suzy throws a rock at a bottle one second before Billy does. Her rock hits the bottle squarely, and it shatters. Billy's rock would have hit the bottle one second later, but instead it sails over the broken shards unimpeded. Unlike the earlier case of preemption, in which Billy drops his rock, the backup process is not derailed until the effect occurs. Up until the moment when the bottle shatters, Billy's rock remains on track to hit the bottle in the (hypothetical) event that Suzy's rock miraculously disappears or swerves off course. Lewis (1986b) called this kind of case "late preemption," although he later (2000) suggested "late cutting" as more appropriate. In this kind of case, the strategy of finding a chain of causal dependence will not work.

Lewis's solution was to consider the class of *intrinsic duplicates* of the process consisting of Suzy's throw, her flying rock, her rock's striking the bottle, and the bottle's shattering. An intrinsic duplicate of this process is a process that takes place in a possible world that has the same laws of nature as our world, and in which the same (or almost the same) intrinsic qualitative properties are instantiated at analogous spatiotemporal regions. An intrinsic duplicate is a little piece of a Humean mosaic that (almost) exactly matches this little piece of the Humean mosaic of the actual world. Lewis (1986b) suggested that causation was an intrinsic manner. This means that if c is a cause of e in our world, c should be a cause of e in every intrinsic duplicate. While the bottle's shattering does not causally depend upon Suzy's throw in the actual world, it will do so in most intrinsic duplicates. In most worlds where throws just like Suzy's occur, there will be no preempted backup. Therefore, Suzy's throw should count as a cause because of the causal dependence that is present in such duplicates. In such a case, Lewis said that the bottle's shattering *quasi-depends* on Suzy's throw.

> **Quasi-dependence:** Let c and e be distinct events that both occur. e quasi-depends upon c iff e causally depends upon c in most intrinsic duplicates of the $c - e$ process.

Causation is then defined as the ancestral of quasi-dependence.

Many of the approaches to causation that have been developed since the time of Lewis's death, including those using structural equations models, have used something like Lewis's idea of quasi-dependence. The idea is that there can be causation without counterfactual dependence so long as there is counterfactual dependence of the effect upon the cause in some appropriate variant of the

actual world. See, for example, Hitchcock (2001), Yablo (2002, 2004), Halpern and Pearl (2005), Hall (2007).

Lewis (1986b) presented these three new proposals separately, and did not formulate a single definition of causation incorporating all three. I think that chancy causation can be readily combined with either of the other two proposals. It is harder to see how to combine causation involving absences with causation as the ancestral of quasi-dependence, a problem that Lewis later seemed to realize.

19.3.3 Causation 2000

When Lewis revised his theory a final time (Lewis 2000, 2004a), he offered several reasons for abandoning the definition in terms of quasi-dependence. The theory had problems dealing with cases of *double prevention*, a kind of case discussed extensively in Hall (2004), a paper that had been circulating since the mid 1990s. Suppose that Billy intercepts a thief intent on stealing Suzy's rock. She later throws her rock at a bottle and shatters it. If Billy had not intervened, Suzy would have had no rock to throw, and the bottle would have remained intact. Billy's intervention is a double preventer of the bottle's shattering: he prevented something – the theft of Suzy's rock – that would have prevented the shattering. A double prevention is a chain involving first prevention, and then causation by omission. Billy's intervention prevented the theft, and the absence of the theft caused the bottle to shatter. If one accepts that there can be causation of and by absences, and that causation is transitive, it seems that one has to accept that double preventers are causes.

Cases of double prevention are doubly problematic for the quasi-dependence account. First, they shed doubt on the guiding intuition that causation is an intrinsic matter. Focus on Billy's interception of the thief and the bottle, and one misses what makes one a cause of the other. It is only because of extrinsic matters, such as Suzy's intentions *vis-à-vis* the rock, that Billy's action has any bearing on the bottle's shattering. Second, there does not appear to be any kind of process connecting Billy's action with the shattered bottle. Neither Billy nor the would-be thief actually touched Suzy, her rock, or the bottle. There doesn't appear to be anything to use as a blueprint for making intrinsic duplicates.

The definition in terms of quasi-dependence also has trouble with a new kind of preemption called *trumping* preemption, introduced by Jonathan Schaffer (2000). Mikey is very gullible. He will do anything that either Billy or Suzy tells him to do. If Billy tells him to jump, he jumps. If Suzy tells him to run, he runs. But if Suzy and Billy give him simultaneous conflicting orders, he always obeys Suzy. It is not that the process connecting Billy's order to Mikey's action is cut short; for example, it is not that Suzy's order initiates a chain of events that shuts off the "obey Billy" module in his brain. If both Billy and Suzy give orders, Mikey hears them both, processes them both, but then only follows Suzy's order. Suzy and Billy both tell Mikey to run, and he runs. Shaffer (2000) argues, and Lewis (2000, 2004a) agrees, that only Suzy's order caused him to run. If Suzy had told him to jump or lie down, Mikey would have done that instead. If Billy had told him to jump or lie down, Mikey still would have run. But Mikey's running quasi-depends upon Billy's order. If we just look at the intrinsic process connecting Billy's order to Mikey's action, it is exactly the same as it would have been if only Billy had given an order. But. if only Billy had given an order, Mikey's action would causally depend upon Billy's order. So quasi-dependence doesn't succeed in distinguishing Suzy's preempting order from Billy's preempted order. (Not all critics agree that only Suzy's order is a cause: see McDermott 2002, Halpern and Pearl 2005, and Hitchcock 2011a.)

Lewis developed his new account through a series of definitions (page references are to the extended version, Lewis 2004a):

> **Alteration:** Let an *alteration* of event *e* be either a very fragile version of *e* or else a very fragile alternative event which may be similar to *e*, but is numerically different from *e*. One alteration of *e* is the very fragile version that actually occurs: the *unaltered* alteration, so to speak. The rest are unactualized. (2004a, 88; here and in other passages I have changed the notation slightly to maintain consistency)

Suppose *e* is an event involving the motion of Billy's car. *e* itself may be fragile or not. Alterations of *e* will be very fragile events such as Billy's car moving at 90 km per hour due east (an event that would not occur if the car were going slightly faster or slower, or in a slightly different direction).

Lewis's key new concept is *influence*:

> **Influence:** Where *c* and *e* are distinct actual events, let us say that *c influences e* iff there is a substantial range $c_1, c_2, \ldots$ of different not-too-distant alterations of *c* (including the actual alteration of *c*) and there is a range $e_1, e_2, \ldots$ of alterations of *e*, at least some of which differ, such that if c_1 had occurred, e_1 would have occurred, and if c_2 had occurred, e_2 would have occurred, and so on. (2004a, 91)

Lewis modified these definitions slightly to allow for alterations and influence involving absences of events. Influence is a special case of causal dependence among families of events (if one loosens the definition of causal dependence in the ways recommended above).

Lewis then defines causation as the ancestral of influence. For convenience, I will allow *c* and *e* to refer to absences as well as events, and say that an absence "occurs" if the corresponding event does not occur.

> **Causation (2000):** Let *c* and *e* be distinct events (or absences) that actually occur. Then *c* is a cause of *e* iff there is a chain of events (or absences) $c = d_1, d_2, \ldots, d_n = e$, all of which occur, such that for each $i = 1, \ldots, n - 1$, d_i influences d_{i+1}.

Before examining this definition in greater detail, let's see how this definition is supposed to handle the trumping cases that brought the quasi-dependence account to grief. Consider a range of alterations of Suzy's order in which she gives different orders: an order to run, and order to jump, and order to lie down, etc. Similarly, consider a range of alterations in which Mikey performs various actions. If Suzy had given a different order, Mikey would have performed a correspondingly different action. Thus Suzy's order to run *influences* Mikey's action of running. By contrast, Billy's order does not influence Mikey's action. If Billy had given a different order, Mikey still would have run.

The definition of influence contains two vague expressions: there must be a "substantial range" of alterations of *c*, and these must be "not-too-distant." Lewis accepts the vagueness, and concludes that influence, and hence causation, is a matter of degree:

> Influence admits of degree in a rough and multi-dimensional way. How many different c_i's are there? How distant are the rest of them from the actual alteration of *C*, and from one another? How much do the e_i's differ from one another: How many different ones are there, and when two of them do differ, how distant . . . are they? (Lewis 2004a, 92)

As we will see with the help of an example, the role of "distance" is different for *c* and for *e*.

Consider a case of late preemption, where Suzy's rock hits the bottle one second before Billy's does. Suzy's throw exerts a great deal of influence on the shattering of the bottle. If she had thrown a little

sooner or a little faster, the bottle would have shattered a little earlier. If she had aimed a little higher or lower, the bottle would have shattered in a slightly different way. Similarly, if her rock had been a little bigger or smaller, or if it had a slightly different shape, the shattering would have been different. Billy's throw exerts *some* influence on the shattering. His rock has a gravitational influence on Suzy's rock and the bottle, his throw will produce little eddies of air, and minor changes in temperature. These will all make some difference for the exact details of the shattering. So different alterations of Billy's throw will lead to different alterations of the shattering. But the changes that these alterations in Billy's throw would make are tiny. So in comparison with Suzy's throw, Billy's throw exerts only a tiny amount of influence on the shattering – perhaps no more than the moon, a nearby tree, a passing truck, and so on. When the influence is this negligible, we tend to ignore it and classify the event in question as a non-cause.

In this example, then, the distance of the alterations of the effect from each other is the main factor that distinguishes Suzy's throw from Billy's throw. Greater distance makes for more influence.

But there is another consideration we must attend to in this example. Suppose Suzy had thrown her rock a little bit later. If she had thrown her rock a fraction of a second later, the bottle would have shattered a fraction of a second later. But if she had thrown her rock more than a second later, the bottle would have shattered exactly one second later, in the manner determined by the particulars of *Billy's* throw. Indeed any alteration of Suzy's throw in which she throws more than a second later will result in the *same* alteration of the shattering. If she throws it two seconds later and a bit higher, or three seconds later and a bit lower, the bottle would have shattered at the same time and in the same manner. Suzy's influence runs out once we alter her throw sufficiently in a particular direction. Similarly, once we consider alterations in which Billy throws more than a second sooner, all of a sudden he begins to exert a great deal of influence over the shattering. If he had thrown two seconds sooner, the bottle would have shattered a second sooner, and in a different manner, than it actually did. And if Billy had thrown two seconds sooner and a little higher, the shattering would have been different again. So perhaps the influence of Billy's throw is not so different from that of Suzy's.

Lewis responds to this problem in a footnote:

> In a context in which we're comparing Billy's throw and Suzy's, such alterations [in which Billy's rock arrives first] should be set aside as "too distant." (Lewis 2004a, n. 23)

This suggests that sensitivity of the effect to alterations of the cause that are very close to the actual alteration is of central importance. When it comes to alterations of the effect, distance from actuality enhances influence; but when it comes to alterations of cause, proximity to actuality enhances influence.

Unfortunately, it is hard to square this treatment of late preemption with Lewis's intended treatment of double prevention. Looking at the definition of influence, we can see that causal dependence between single events is a potential case of influence. Lewis writes:

> [W]e do not require a many–many mapping; the simplest sort of whether–whether dependence, with only two different alterations of *e*, still qualifies as one sort of pattern of influence. (Lewis 2004a, 91–2)

By "whether–whether dependence," Lewis means dependence of whether or not *e* occurs upon whether or not *c* occurs: the kind of dependence constituting causal dependence of *e* on *c*. In cases of double prevention, we often have only this kind of whether–whether dependence. Consider the case where Billy intercepts a thief intent on stealing Suzy's rock, which Suzy later uses to smash

a bottle. If Billy had not intercepted the thief, the bottle would not have shattered: we have whether–whether dependence. But if Billy had intercepted the thief a little bit earlier or a little bit later, if he had intercepted the thief in a different manner (perhaps coming at the thief from a different angle), there would have been no corresponding changes in the time or manner of the bottle's shattering. We have only whether–whether dependence.

This kind of whether–whether dependence scores very poorly along most of Lewis's dimensions of influence. There are only two alterations of the cause – the event occurs or it doesn't. The non-actual alteration is maximally distant – it requires that the event not occur at all. And there are only two alterations of the effect – again, it either happens or it doesn't. At least these two alterations are distant, and that enhances influence. Consider especially the second dimension, how distant the alterations of the cause are. In our case of late preemption, an alteration of Billy's throw where he throws more than a second sooner was considered "too distant." In our case of double prevention, why is the alteration in which Billy does not intercept the thief at all not "too distant"? Moreover, Suzy's throw outscores Billy's intervention along three of the four dimensions of influence, and equals it along the fourth. There are many different alterations of Suzy's throw that would lead to different alterations of the shattering, and some are very close to Suzy's actual throw. Moreover, the alteration in which the bottle doesn't shatter at all is still included among the alterations that can result from alterations in Suzy's throw. Why doesn't the influence of Suzy's throw swamp the influence of Billy's intervention, just as Suzy's throw swamped the influence of Billy's throw in our case of late preemption?

19.4 Applications

Over the course of developing his three analyses of causation, Lewis introduced a number of subsidiary concepts: causal dependence among families of events, causal dependence among individual events, chancy causal dependence, quasi-dependence, influence. What I will now argue is that when Lewis provided causal accounts of further concepts, he did not make use of "causation" as he analyzed it. Rather, he made use of one of these other concepts, or else fashioned new concepts out of the same basic parts.

19.4.1 Causal Decision Theory

Lewis (1981b) advocated a version of causal decision theory, in which a rational agent evaluates her potential actions according to what they might cause. Let C be the credence function of a rational agent, and let V be her utility function. Suppose she must choose from among actions $\{A_1, A_2, \ldots, A_m\}$. Finally, let $\{K_1, K_2, \ldots, K_n\}$ be the set of *dependency hypotheses* (of which more anon) to which the agent assigns non-zero credence. Then the causal expected utility of act A_i is:

$$U(A_i) = \Sigma_j C(K_j)V(A_iK_j) \tag{19.1}$$

Causal decision theory then recommends that the agent choose the act with the highest causal expected utility.

The key causal notion here is that of a dependency hypothesis. Lewis says that a dependency hypothesis is "a maximally specific proposition about how the things [the agent] cares about do and do not depend causally on his present actions" (Lewis 1981b, 313; here and elsewhere, page references are to the version reprinted in Lewis 1986a). This suggests that a dependency hypothesis is a hypothesis about the causal dependence of one family of events upon another. That is almost right.

Let $\{S_1, S_2, \ldots, S_r\}$ be a partition of outcomes that is fine enough to capture the differences that the agent cares about (what Lewis 1981b, 317 calls a *rich* partition). For example, if the agent is playing a game with cash prizes, and cares only about money (and not prestige, winning for its own sake, etc.) then $\{S_1, S_2, \ldots, S_n\}$ could be a partition into prize amounts. Then one kind of dependency hypothesis would be as follows:

> **Dependency hypothesis (Gibbard-Harper):** A Gibbard-Harper dependency hypothesis is a conjunction of counterfactuals of the form $A_i \rightarrow S_j$, with one such conjunct for each possible action.

Such a conjunction would describe a pattern of causal dependence (or a lack of causal dependence) of $\{S_1, S_2, \ldots, S_n\}$ on $\{A_1, A_2, \ldots, A_m\}$. Lewis (1981b, 328) shows that if all of an agent's dependency hypotheses have this form, his causal decision reduces to the version of causal decision theory advocated by Gibbard and Harper (1978).

More generally, Lewis takes a dependency hypothesis to have the following form:

> **Dependency hypothesis:** A dependency hypothesis is a conjunction of counterfactuals of the following form: '$A_i \rightarrow$ the objective chance of S_j is p'; with one conjunct for each combination of A_i and S_j.

If we expand the definition of causal dependence among families of events to allow for chancy dependence (in the same way that Lewis 1986b expanded the definition of causal dependence among single events to allow for chancy dependence), then a dependency hypothesis describes a pattern of causal dependence (or a lack of causal dependence) of $\{S_1, S_2, \ldots, S_r\}$ on $\{A_1, A_2, \ldots, A_m\}$.

The point I want to stress is that the key concept is not *causation*, but rather a probabilistic extension of causal dependence among families of events. Indeed, issues that seemed central to defining causation are *irrelevant* to decision. Suppose that Billy is listening to his favorite song on his stereo.[6] Consider the various buttons and knobs on the stereo. By pressing the "power" button, he can turn the music on or off, but can't exercise any other control over the sound. By turning the "volume" knob, he can make the sound louder or quieter, and even make it so quiet as to be inaudible. By turning the "treble" and "bass" knobs he can change the balance of high and low frequencies in the sound, but he cannot use these to eliminate the sound altogether. The power currently on, the other knobs are at specific levels, and audible music is coming from the speakers. Which knob settings are *causes* of the sound? In Lewis's original theory (Lewis 1973a), the power button's being set to "on" is a cause. The settings of the base and treble knobs are not causes. Whether the setting of the volume knob is a cause will (arguably) depend upon exactly which event we have in mind – a very fragile event in which the volume must be set to its precise level, or a more robust event in which the volume must only be set above some threshold. In Lewis's final theory (Lewis 2000, 2004a), the volume knob exerts the greatest influence over the sound, so it will be a cause. The bass and treble knobs exert considerable influence as well, so they would probably qualify as causes of the sound. The case of the power button is more problematic, since it exhibits only the kind of "whether–whether" dependence that gets the lowest score on almost all dimensions of influence.

But if Billy must decide how to adjust the stereo, there is no need to settle the question of which knob settings are "causes" of the sound. Rather, his utilities will determine which adjustments he should make. If Suzy is napping, and Billy wishes not to wake her, he should turn the power off, or else turn the volume down; adjusting the bass and treble won't help. If he wants to be able to hear the stereo from the kitchen while he prepares dinner, only the volume knob will help him. And if he wants to impress Suzy, an audiophile, with the quality of the sound, he will want to able to fine tune

the sound with the base and treble knobs. The power button wouldn't help him: even the cheapest stereo systems can be turned off.

Indeed, it is not clear to me what the point would even be of saying that some of these knob settings are *causes* of the sound, while others are not. The sound causally depends upon the settings of the different knobs in different ways. Which kinds of causal dependence matter will depend upon our purposes.

Now suppose that Billy and Suzy are back to their old tricks, throwing rocks at a bottle. Suzy wants the bottle to shatter, and only cares about whether the bottle shatters. She knows that if she throws, her rock will shatter the bottle; if she doesn't throw the rock, Billy will throw his, and the bottle will shatter. Whether or not the bottle shatters does not depend upon whether or not she throws. Should she throw her rock or not? It seems obvious that she should be indifferent between throwing and not throwing. What matters here is the lack of causal dependence. There is no need to add any bells and whistles to the basic concept of causal dependence in order to deal with preemption. If Suzy throws, her throw will preempt Billy's, and her throw will cause the bottle to shatter. But none of this matters to her decision; all that matters is whether or not shattering causally depends upon her throw.

This is rather strange. Concerns over preemption were a primary driver in motivating revisions to Lewis's account of causation, and they have been ubiquitous in the causation literature. But these concerns are completely irrelevant to decision, which seems like a paradigmatically causal concept. This suggests that the centrality of preemption in discussions of causation is at least partially misplaced.[7]

19.4.2 Dispositions

A simple conditional analysis of dispositions along the following lines was once widely accepted:

> Something x is disposed at time t to give response r to stimulus s iff, if x were to undergo stimulus s at time t, x would give response r. (Lewis 1997, 133; here and elsewhere, page references are to the version reprinted in Lewis 1999)

Martin (1994) presents counterexamples to this analysis involving dispositions, and lacks of disposition, that are *finkish*. Suzy has been studying the dark arts. This has caused some strange changes in her personality. She has taken an intense liking to a certain bottle, and an intense dislike to a certain rock. She has resolved that if the bottle is ever struck, she will instantly cast a spell transforming it from glass to solid stone, thus preventing it from breaking. Similarly, if the rock is ever struck, she will cast a spell transforming it from solid rock to delicate crystal, allowing it to break from the striking. At time t, Suzy has yet to act on either commitment. We want to say at time t that the bottle is fragile, that is, disposed to break if struck, while the rock is not. But the conditional analysis fails for both objects. The bottle's disposition to break is *finkish*, it will disappear if the relevant stimulus conditions occur. Similarly, the rock's lack of fragility is finkish. A disposition of the bottle or the rock must be something intrinsic to the bottle or the rock, and not contingent upon external intervention, such as Suzy's spells.

Lewis offers an alternative analysis:

> **Disposition (Lewis):** Something x is disposed at time t to give response r to stimulus s iff, for some intrinsic property B that x has at t, for time t' after t, if x were to undergo stimulus s at time t and retain property B until t', s and x's having of B would jointly be an x-complete cause of x's giving response r. (Lewis 1997, 149)

I will not attempt to evaluate this new analysis. But I want to point out that once again, Lewis has not made use of any of his own analyses of causation. Instead, he has crafted a new causal concept, that of an *x*-complete cause. An *x*-complete cause is "a cause complete in so far as havings of properties intrinsic to *x* are concerned" (Lewis 1997, 149). I take this to mean:

> ***x*-complete cause:** *s* and *x*'s having of *B* form an *x*-complete cause of *x*'s giving response *r* iff: (i) *s* is a cause of *x*'s giving response *r*; (ii) *x*'s having *B* is a cause of response *r*; and (iii) if any of *x*'s other intrinsic properties (besides *B*) had been different, *x* still would have given response *r*.

Clauses (i) and (ii) appeal to the ordinary notion of causation. But clause (iii) adds a requirement that a certain relation of causal dependence not hold. This concept of an *x*-complete cause is characterized using the same basic tools that were used in Lewis's analyses of causation, but once again, it is not the concept of causation itself that is needed.

I think this is particularly surprising in the case of dispositions. The problem of finkish dispositions is pretty clearly a problem involving preemption. Hence, one would think that an analysis of causation that had been engineered to handle cases of preemption would be well suited to addressing the problem of finkish dispositions. More specifically, it seems to me that the concept of *quasi-dependence* might be helpful. Recall that Lewis abandoned his analysis of causation in terms of quasi-dependence in part because he rejected the intuition that causation was an intrinsic matter. But in the case of dispositions, we want the relationship between the stimulus and the response to be mediated by the intrinsic properties of the object in question. This suggests to me that the following might be a promising start toward an analysis of dispositions:

> **Disposition (in the spirit of Lewis):** Something *x* is disposed at time *t* to give response *r* to stimulus *s* iff most of *x's* intrinsic duplicates *y* are such that if *y* were to undergo stimulus *s* at time *t*, *y* would give response *r*.

This is equivalent to:

> **Disposition (in the spirit of Lewis):** Something *x* is disposed at time *t* to give response *r* to stimulus *s* iff its failure to give response *r* quasi-depends upon the absence of stimulus *s* at time *t*.

I do not know if this analysis will succeed, but it seems prima facie promising. The general point is that thinking more carefully about the connections between the problems we encounter and the tools we use when trying to analyze causation, and the problems we encounter and the tools we use when trying to provide causal analyses of other concepts, is likely to be fruitful.

19.5 Conclusion

Figure 19.2 shows a family tree of concepts, with causal dependence among families of variables at the trunk. These are (some of) the concepts developed by Lewis in his attempts to analyze causation, and in his attempts to provide causal accounts of rational decision and dispositions. If you zoom a microscope on "causation" in Figure 19.1, this is what you would see. My own terminological preference is to use "causation" to refer to the whole family of concepts descending from causal dependence, rather than to one specific concept in the tree. This fits better with Lewis's use of expressions like "causal dependence" and "causal decision theory." Following the terminology of Pearl (2000), I prefer to call the specific relation that Lewis repeatedly tried to analyze "actual causation." Each of

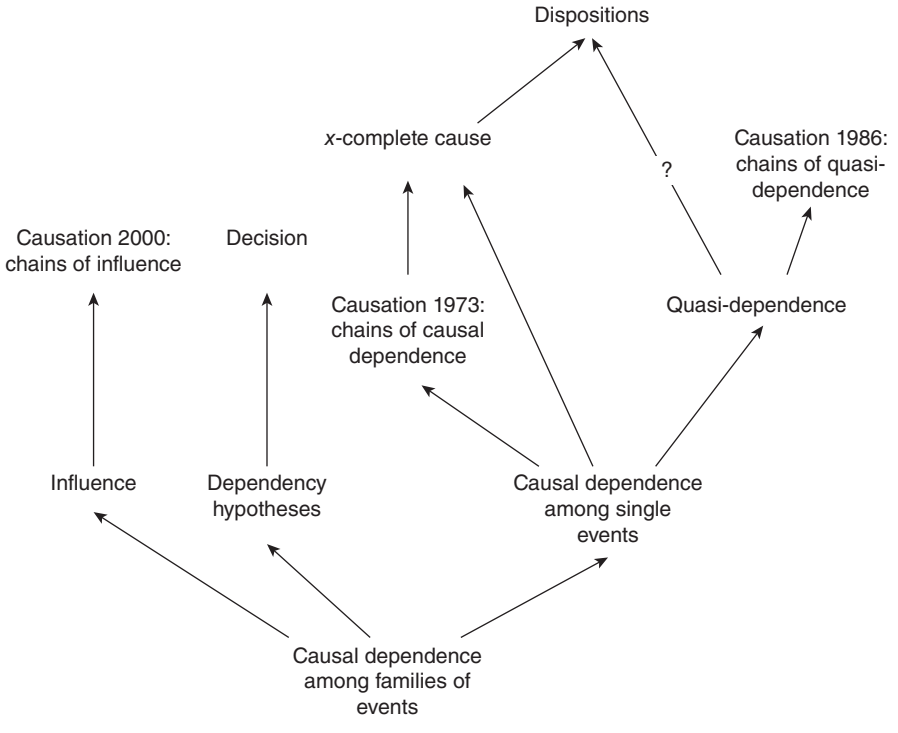

Figure 19.2 Interrelationships among concepts introduced in Lewis's analyses of causation, rational decision, and disposition. The arrow marked with a question mark corresponds to my own proposal. Created by Christopher Hitchcock.

the concepts in the tree has an internal structure to it, and thinking about that structure in terms of non-backtracking counterfactuals is extraordinarily fruitful. Understanding the internal structure of a concept, as well as its relation to other concepts, will help us to better predict which concepts are likely to be well suited for which purposes. For example, instead of worrying about whether causation is transitive or not, it might make more sense to ask which applications require a transitive causal relation, and which do not.

Thinking in terms of a family tree of causal concepts also leads to a number of important new questions. For example, if the specific concept of actual causation is not well suited for analyses of rational decision and disposition, why do we seem to place such great importance upon it? What *is* it well suited for? It seems to be well suited for explanation, and for notions of moral and legal responsibility (see e.g. Moore 2009). But *why* is it well suited for these applications? What is it about the *structure* of the concept that makes it well suited for these applications, but not others? It will not do to say it is because explanation, and moral and legal responsibility, all involve *causation*. So do decision and disposition. Answering these kinds of questions will tell us a great deal, both about causation, and about related concepts that have a causal component. And it is in the service of answering these kinds of question that analyses such as those offered by Lewis, even imperfect and incomplete ones, may have their greatest value.

Acknowledgments

For comments and suggestions, I would like to thank Adam Elga, Elizabeth Harman, Frank Jackson, Boris Kment, Stephanie Lewis, Laurie Paul, Gideon Rosen, and audience members at Princeton University.

Notes

1 I am using "concept" as a catch-all term for the kind of thing philosophers like to give analyses of. It is not intended to designate a mental representation.
2 I believe that this canonical example, and its canonical characters, originally appeared in an early version of Hall 2004, although they appear in print prior to 2004, e.g. in Lewis 2000.
3 An exception, I think, is Lewis's account of causal explanation in Lewis 1986d.
4 Definitions that appear without a citation to Lewis are in my own wording.
5 See Hitchcock 2004 for more discussion of this kind of case.
6 This example is similar to one I give in Hitchcock 2011b.
7 For further discussion of this issue, see Hitchcock forthcoming.

References

Collins, J., Hall, N. & Paul, L. (Eds.) (2004). *Causation and counterfactuals*. Cambridge, MA: MIT Press.
Gibbard, A. & Harper, W. (1978). Counterfactuals and two kinds of expected utility. In C. Hooker, J. Leach & E. McClennen (Eds.), *Foundations and applications of decision theory*, pp. 125–62. Dordrecht: Reidel.
Hall, N. (2004). Two concepts of causation. In Collins, Hall & Paul (2004), pp. 225–76.
Hall, N. (2007). Structural equations and causation. *Philosophical Studies* 132, 109–36.
Halpern, J. & Pearl, J. (2005). Causes and explanations: a structural-model approach – part I: causes. *British Journal for the Philosophy of Science*, 56, 843–87.
Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy*, 98, 273–99.
Hitchcock, C. (2004). Do all and only causes raise the probabilities of effects? In Collins, Hall & Paul (2004), pp. 403–17.
Hitchcock, C. (2011a). The metaphysical bases of liability: commentary on Michael Moore's *Causation and responsibility*. *Rutgers Law Journal*, 42, 377–404.
Hitchcock, C. (2011b). Trumping and contrastive causation. *Synthese*, 181, 227–40.
Hitchcock, C. (forthcoming). What is the "cause" in causal decision theory? *Erkenntnis*.
Lewis, D. (1973a). Causation. *Journal of Philosophy*, 70, 556–7. Reprinted in Lewis (1986a), pp. 159–72.
Lewis, D. (1973b). *Counterfactuals*. Cambridge, MA: Harvard University Press.
Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, 13, 455–76. Reprinted in Lewis (1986a), pp. 32–52.
Lewis, D. (1980). A subjectivist's guide to objective chance. In R. Jeffrey (Ed.), *Studies in inductive logic and probability*, Vol. II, pp. 263–94. Berkeley: University of California Press. Reprinted in Lewis (1986a), pp. 83–113.
Lewis, D. (1981a). Kausalität: epilog 1978. In G. Posch (Ed.), *Kausalität – neue texte*. Stuttgart: Reclam.
Lewis, D. (1981b). Causal decision theory. *Australasian Journal of Philosophy*, 59, 5–30. Reprinted in Lewis (1986a), pp. 305–37.
Lewis, D. (1986a). *Philosophical papers*, Vol. II. Oxford: Oxford University Press.
Lewis, D. (1986b). Postscripts to "Causation." In Lewis (1986a), pp. 172–213.
Lewis, D. (1986c). *Events*. In Lewis (1986a), pp. 241–70.
Lewis, D. (1986d). Causal explanation. In Lewis (1986a), pp. 214–40.

Lewis, D. (1997). Finkish dispositions. *Philosophical Quarterly*, 47, 143–58. Reprinted in Lewis (1999), pp. 133–51.
Lewis, D. (1999). *Papers in metaphysics and epistemology*. Cambridge: Cambridge University Press.
Lewis, D. (2000). Causation as influence. *Journal of Philosophy*, 97, 182–97.
Lewis, D. (2004a). Causation as influence, expanded version. In Collins, Hall & Paul (2004), pp. 75–106.
Lewis, D. (2004b). Void and object. In Collins, Hall & Paul (2004), pp. 277–90.
Martin, C.B. (1994). Dispositions and conditionals. *Philosophical Quarterly*, 44, 1–8.
McDermott, M. (1995). Redundant causation. *British Journal for the Philosophy of Science*, 46, 523–44.
McDermott, M. (2002). Causation: influence vs. sufficiency. *Journal of Philosophy*, 99, 84–101.
Moore, M. (2009). *Causation and responsibility: an essay in law, morals, and metaphysics*. Oxford: Oxford University Press.
Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge: Cambridge University Press.
Schaffer, J. (2000). Trumping preemption. *Journal of Philosophy*, 97, 165–81. Reprinted in Collins, Hall & Paul (2004), pp. 59–73.
Yablo, S. (2002). De facto dependence. *Journal of Philosophy*, 99, 130–48.
Yablo, S. (2004). Advertisement for a sketch of an outline of a prototheory of causation. In Collins, Hall & Paul (2004), pp. 119–37.

Part IV

Language and Logic

20

David Lewis on Convention

ERNIE LEPORE AND MATTHEW STONE

David Lewis's landmark *Convention* starts its exploration of the notion of a convention with a brilliant insight: we need a distinctive social competence to solve coordination problems. Convention, for Lewis, is the canonical form that this social competence takes when it is grounded in agents' knowledge and experience of one another's self-consciously flexible behavior. Lewis meant for his theory to describe a wide range of cultural devices we use to act together effectively; but he was particularly concerned in applying this notion to make sense of our knowledge of meaning. In this chapter, we give an overview of Lewis's theory of convention, and explore its implications for linguistic theory, and especially for problems at the interface of the semantics and pragmatics of natural language. In section 20.1, we discuss Lewis's understanding of coordination problems, emphasizing how coordination allows for a uniform characterization of practical activity and of signaling in communication. In section 20.2, we introduce Lewis's account of convention and show how he uses it to make sense of the idea that a linguistic expression can come to be associated with its meaning by a convention. Lewis's account has come in for a lot of criticism, and we close in section 20.3 by addressing some of the key difficulties in thinking of meaning as conventional in Lewis's sense. The critical literature on Lewis's account of convention is much wider than we can fully survey in this chapter, and so we recommend for a discussion of convention as a more general phenomenon Rescorla (2011).

20.1 Coordination in the Social World

Lewis regarded *coordination problems* as providing a distinctive insight into the social world. Coordination problems name a subset of strategic situations where agents must choose actions that agree with one another to achieve mutually acceptable outcomes. In coordination problems, agents have to make matching choices from among multiple candidate strategies, but there's no intrinsic reason to prefer one over another. So, agents must act solely based on their expectations about one another. Coordination problems, we shall see, require a distinctive kind of reasoning and knowledge. Uniquely, in coordination problems, agents' expectations about one another are both decisive in settling what

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

they should do and independent of the practical logic of their situation. The phenomenon of coordination thus brings powerful constraints to anchor our thinking about social competence.

The examples that Lewis gives show the scope of this phenomenon. One example is holding a meeting. If we're arranging as a group where to hold a meeting, we could hold it at your place, in Princeton perhaps. Or we could hold it at our place, in New Brunswick. Considering our joint effort, it's equally efficient for us to meet in either place. We have multiple options, and it's up to us what we do. Practical logic doesn't dictate where we go. Nevertheless, what we do has to match. If half of us end up in Princeton and half in New Brunswick, we won't have a meeting at all. You share with us our interest in going to the same place. So, you should go where you expect we will be waiting to meet you, and we will do likewise. Our expectations about one another are decisive.

Once one recognizes coordination as a phenomenon, examples proliferate. A vivid and dramatic case is the choice whether to drive on the right side of the road, as Americans and Continental Europeans do, or whether to drive on the left side of the road, as they do in the UK and Japan. In choosing to go right, we are acting on our expectations of how others will choose to go. We think they will also go right. If we thought they would all go left, then we would have gone left as well; we all wish to avoid congestion or collision. The coordinated expectations that send us all to the right – as in the USA – do not make reference to anything in the world that settles preferences or strategies for us. Going left is just as good. Either strategy serves its participants equally well, just so long as all of the participants agree on the same solution.

Lewis also considers resuming a dropped phone call between two parties. It's in their interest to pick up the conversation and continue where they left off. But if both dial immediately, the connection fails. Exactly one party must attempt to initiate it. You might have a variety of ways of stating a rule or policy for settling who does what. According to Lewis, the operative one in Oberlin, back when he was growing up, was that the person who dialed the original call was the one who must redial when the call is dropped. The person who received the original call must wait. We have acquaintances who still follow this strategy today, in the days of cell phones, where even the receiver has a button to redial the last incoming caller. The case illustrates a point Lewis emphasizes: to solve coordination problems, agents must do their part; it may not be enough for them simply to act the same way.

The point extends to coordination among multiple agents. Lewis considers a group that must carry out a search together. They maximize their success when each person in the party stakes out a different area to search. When everyone searches the same area, territory remains unexplored. Their search may fail. So, again, everybody has to choose his action depending on what everybody else chooses.

It may be that the coordination activity is something hard to describe, implicit or continuous. Lewis uses the example of two people in a rowboat from Hume (1740, §3.2.2). The two people have to row in concert in time for their strokes to be effective and the motion of the boat smooth. But it doesn't particularly matter how fast they row within certain limits; they can adjust the force they use so that it's a reasonable pace no matter how fast their interval is. Normally, they can get into sync without ever making an explicit agreement or even discussing their pace, simply by making small adjustments to one another. In the end, they wind up not only matching their motions, but also matching their expectations of one another. Each chooses to row at a certain speed because he expects the other to do so.

Lewis develops a formal approach to coordination and communication by drawing on game theoretic ideas from von Neumann and Morgenstern (1947) and Schelling (1960). We review his formalism briefly here; we recommend Clark (2011) for a gentle introduction to game-theoretic analysis of coordination and communication; a broader technical overview of the subject from a philosophical perspective can be found in Skyrms (2010).

	c_1	c_2
r_1	1, 1	0, 0
r_2	0, 0	1, 1

Diagram 20.1

A game is a formal way of describing a situation in which multiple agents interact, make choices and achieve outcomes. Games generalize the decision-theoretic models that characterize the choices of individual agents acting alone. In decision-theoretic models, agents make choices by anticipating the expected outcome of their actions in virtue of the laws of nature. In games, multiple agents must make choices, perhaps without being able to observe one another's actions. To anticipate the outcomes of actions in games, agents must reason not only about the natural course of events but also about the choices other agents will make. This gives an inherently strategic character to ideal reasoning in game-theoretic models.

The simplest formalization of games is simply to map out the choices that are available to each agent and the payoffs that result for each combination of actions on the part of the players. This is known as the *extensive form* of a game. The game in Diagram 20.1 typifies the situations that Lewis associates with coordination.

It describes an interaction between two agents, the *row player* and the *column player*. They each select one of two possible actions; then they get a payoff. Coordination is required here because the row player and the column player both get a good payoff if they choose their first action, and they both get a good payoff if they choose their second action. But they get bad outcomes if their choices don't match. Their preferences are aligned in this case, and their preferences dictate that they make matching choices, but their preferences don't dictate which choice they will make or must make or should make. Either matching strategy is just as good.

What this shows is that the intuitive notion of coordination that Lewis has identified in his vignettes from everyday life has a mathematical basis. Formalizing coordination in terms of games highlights what's distinctively social about coordination. The right action is inherently underdetermined by agents' interests and the practical logic of the situation – that is, the way actions lead to outcomes. Agents' expectations about one another are decisive and, in a certain sense, arbitrary.

We can make this more precise by appealing to the natural game-theoretic account of rational choice. Consider an agent in a game. If she chooses her actions rationally, based on their expected outcomes, she needs to consider her expectations about the other agents' actions, because those actions are going to determine her payoffs in part. If the row player knows the column player will do c_1, she ought to do r_1. If she knows that the column player will do c_2, she ought to do r_2. But the same is true of the column player. If he knows that the row player will do r_1, he ought to do c_1. If he knows that the row player will do r_2, he ought to do c_2. Thus, a good way to think about a satisfactory outcome in a game like this – what is it that agents should aim for, and what is it that their behavior should achieve – is as a *Nash equilibrium* (Nash 1950). A Nash equilibrium is a set of strategies, chosen simultaneously by all the players, in such a way that each player achieves the best outcome possible *given* the strategies that all the other players are following. As we have seen, in (1), both c_1–r_1 and c_2–r_2 are equilibria. The formalism captures the intuitive arbitrariness of agents' behavior by the fact that the game is associated with these two different equlibria.

Lewis defines coordination problems as situations of interdependent decision by at least two agents where coincidence of interest predominates and where there are at least two equilibria. In these cases, agents that are coordinating successfully must have additional knowledge of one another – some

social knowledge or competence – that allows them to find the particular equilibrium they use as opposed to the others.

For Lewis, *communication* involves particular kinds of coordination problems. The relationship between communication and coordination is clearest in simple idealized situations known as sender–receiver games. In these situations, one person produces a *signal* knowing privately some aspect of the state of the world. The other person, the recipient, acts having seen the signal. What we are calling signals in these games are just arbitrary actions with no real-world effects that agents intrinsically prefer. Signals get their power through the value of information, as the receiver makes better decisions by recognizing the strategy the sender uses in presenting the signal they do.

The story of Paul Revere's ride, as portrayed in Longfellow's historically embellished 1860 poem, provides the canonical example. Paul Revere and the sexton of Boston's Old North Church must coordinate the militia's defenses against British troops. Revere must determine where the defenses will be mounted, but only the sexton has the key information about how the attack will proceed. The sexton hangs a pattern of lanterns in the church belfry, which can be seen from far away, as a warning signal. He is to use one light if the British are coming by land, and two lights if by sea. In themselves, the lanterns are little more than decorative. But, because of the sexton's plan, the pattern of lights is contingent on what the British are doing, and Revere knows this. So, Revere orchestrates his pattern of defense contingent on the lights he sees. If he sees one lantern hanging in the belfry, he prepares against a land attack; if he sees two, he prepares against a sea attack. The joint strategy of Revere and the sexton allows them to respond effectively no matter how the attack comes. In fact, it is another equilibrium. The sexton has no reason to change his strategy, given what Revere will do, and Revere has no reason to change his strategy, given what the sexton will do.

We can schematize key elements of the Revere story in a formal game. The formalism highlights key features of Lewis's understanding of signaling. For Lewis, coordination is fundamental to signaling. And, for Lewis, what we call the meanings of signals are just agents' expectations about when signals are to be used and what responses they will elicit. Say, the sexton has two lanterns, and will hang them differently depending on his knowledge of the British plans. That makes for two possibilities: one if by land, two if by sea (L_1S_2); and one if by sea, two if by land (S_1L_2). Let's assume that Revere, too, will act differently depending on what he sees: that makes for two possibilities: land defense for one light, sea for two (L_1S_2); and sea defense for one light, land for two (S_1L_2). Here are the payoffs in Diagram 20.2.

So, Revere and the sexton reach an equilibrium when both play L_1S_2. And they also reach one when both play S_1L_2. Either works, but they must coordinate. It is their mutual expectations that are decisive in what they actually do. Mathematically, we see the commonality with straightforward coordination problems, like what side of the road to drive on. The fact that the sexton's actions here have no intrinsic effects gives an added nuance to the case. What Revere and the sexton's mutual expectations amount to here is an *interpretation* for the actions of the sexton as signals of what the British are doing.

Although this abstraction allows Lewis to characterize the *meanings* of signals directly in terms of agents' information, choices, and outcomes, the characterization of meaning that results is very coarse. A signal that indicates that a proposition P is true, and that therefore prompts the recipient

	L_1S_2	S_1L_2
L_1S_2	1, 1	0, 0
S_1L_2	0, 0	1, 1

Diagram 20.2

to choose an action *A*, means just as much *P* as it does "Do *A*." In the Revere story, hanging a single lantern in the belfry could mean the British are coming by land, but it could also be an instruction to prepare the land defenses, or it could be both. All the theory allows us to do is to characterize the probabilistic inference that comes from combining the expectation the sexton will act according to a certain conditional policy and the observation of a specific instance of that policy. The information carried by signals is even further removed from their intuitive meanings in cases where agents' interests do not entirely coincide, so agents have interests in hiding information as well as sharing it, may send unreliable signals and may interpret signals as unreliable. (This has proved problematic not only for human communication but also for understanding the meaning of animal signaling; see Lachmann, Számadó & Bergstrom 2001, for example.) We'll return to this weakness of Lewis's approach in section 20.3.

20.2 Convention

Coordination is a powerful concept for theorizing about social competence. Suppose we observe agents reliably solving a particular coordination problem according to a particular joint strategy. Then they *must* share knowledge or skills that let them choose this pattern over its alternatives. Here's *the* place where agents' knowledge of one another and alignment with one another seems to be doing special work in their abilities to interact – over and above what they know about rationality and over and above what they know about how the world works.

In principle, the mechanisms behind successful coordination could take many forms. Coordination could be the result of agents' innate architecture. This seems likely to be the case, for example, for the famous waggle dance of honeybees (von Frisch 1967). Bees signal the distance and direction of sources of nectar by particular patterns of movement. Other bees can interpret these movements, so they can retrieve the nectar in turn. These movements, like the sexton's signal in the story of Paul Revere, have neither intrinsic meaning nor substantive practical effects. They get their meaning from the strategies of the bees, and in particular from the match between the circumstances in which bees that have found nectar produce the signals and the behaviors with which bees in the audience respond.

If bees had evolved differently, the waggle dance might very well have involved a different set of joint strategies. Indeed, there is some evidence that different species of bees do produce and interpret their waggle dances differently (Su *et al.* 2008). So, the formalism of coordination provides a useful construct for characterizing the social competence of the bees. On the other hand, we needn't suppose that individual bees are rational agents solving this coordination problem on the basis of expectations about one another. Individual bees may have little alternative but to act as they do. Their social competence may be hardwired, or perhaps may allow for only very limited kinds of learning or modulation from interactions with other bees.

Indeed, when biologists approach the evolution of social behavior, they often start from the modeling assumption that agents' choices are simply dictated by their genes. Each individual acts not based on rational calculations but based on a predefined set of rules. However, agents act repeatedly; they have lots of chances to interact with one another across lots of situations. Agents that coordinate often in these interactions are more successful, and thus are more likely to propagate their patterns of choices to their descendants. Under these assumptions, populations will converge to fixed points where the distribution of strategies is under no pressure to change. Those fixed points turn out to be a subset of the familiar Nash equilibria, so-called *evolutionarily stable strategies* (Maynard Smith and Price 1973). The analysis confirms the relevance of the game-theoretic analysis for the theory of social competence, even if agents' individual decisions are heuristic or even fixed in advance.

Here's a somewhat different kind of case. Even if agents have choices to make, reliable coordination could simply be the result of psychological mechanisms that make particular actions salient. In this case, agents don't need to have specific knowledge or expectations about one another to make their choices. They simply do what seems natural to them. Schelling (1960) showed that, in many decision problems, one course of action does seem the simplest or the most natural. It naturally springs to mind or attracts the attention. He called such strategies *focal points*. In many cases, even in novel coordination problems with strangers, people can identify aligned focal points, and so work out how to coordinate successfully much more frequently than one would expect by chance.

However, not all cases of coordination can be explained by such mechanisms of architecture and alignment. Coordination is sometimes a matter of learned expectations about agents' strategies. These cases, for Lewis, are cases of *convention*. A convention, for Lewis, involves the deliberative choice of one regularity R over another regularity R' in a recurrent situation S, in virtue of expectations shared among a population P. In the ideal case (which Lewis allows to be satisfied only to a limited degree), a regularity R in behavior is a convention for P if it's common knowledge among the members of P that they follow R, that they expect one another to follow R, that R is a solution to the coordination problem they face in S, and that there is another solution R' that they could have conformed to instead of R (Lewis 1969, 76).

Note that Lewis's account of conventions does not seem to fit cases of alignment based on hardwired strategies or innate focal points. In these cases the agents' *preferences* in situation S might be satisfied just as well by a regularity other than the one they exhibit; the agents might even know this. But it wouldn't be *possible* for them to conform to any other regularity than the one they actually exhibit. Only if agents have somehow learned or negotiated strategy R does it seem they really could have exhibited some other strategy R'.

Drawing on this framework, Lewis proposes that meaning in language is a matter of convention. His understanding generalizes the simple signaling conventions of the Revere story, where one agent produces an action conventionally under certain circumstances, and other agents respond to that action conventionally in specified ways. For Lewis, those who make utterances, conventionally do so only under particular circumstances: namely, when those utterances are true. Conversely, those who interpret utterances, conventionally act as though the world is a certain way: namely, as though the utterances are true. Such matching strategies on the part of speakers and their audiences constitute equilibria, Lewis suggests. The group prefers the outcomes that follow from effective communication. And there are many such equilibria, as the diversity of human languages attests.

Lewis's definition allows that conventions could be instituted explicitly. Agents might agree to defer to an authority that establishes how they should act together, on the model of the Académie Française. Agents might continue to defer to conventions that date from a particular agreement they regard as binding, like the Geneva Conventions. But Russell (1921) and Quine in the foreword to Lewis (1969) observed that any conventions of language could not all have such explicit origins, on pain of regress. Consider making the first such agreement. The parties would need a language to carry out their negotiations. But since this is the first agreement there could not be any conventions of language yet.

However, Lewis's notion of convention allows for conventions that evolve gradually and tacitly as a side-effect of agents' independent need to make their way in the social world. The first time we face a coordination problem, we have no choice but to do something. So, we do whatever occurs to us. But salience gives us a certain probability of success. It may be that what occurs to us matches what others do, because of our natural affinities. Successful coordination creates a precedent that we can rely on in future interactions. Eventually we have enough confidence from our successful experience coordinating that we have developed the mutual expectations that constitute a convention. As long as agents are able to coordinate in particular kinds of ways, then the conventions are implicit in the

choices they make and the knowledge they have of one another. Because Lewis's definition does not call for explicit agreement in cases of convention, his definition escapes the regress objection.

20.3 Conventions of Meaning in Critical Perspective

Lewis's understanding of convention provides a beautiful way of relating social competence to underlying experience, knowledge and rationality. Lewis's theory of the conventions of language is a brilliant application of the ideas in an attempt to reduce knowledge of language in general, and knowledge of meaning in particular, to our expectations about one another. Moreover, Lewis's account has much in common with H.P. Grice's general view of meaning as a product of interlocutors' collaborative intentions (Grice 1957). To coordinate by convention entails recognizing the intentions of other agents. In particular, if meaning amounts to some conventional regularity R, as Lewis suggests, then in each utterance, the speaker A intends what he does to exemplify R, and his audience B knows this. Moreover, A and B both expect everyone to do their part in R, and do their part conditional on the others doing so. By Lewis's definition of convention, if B thought A intended to participate in R′ and expected B to do so as well, B would participate in R′ instead. Another way to say this is that if B recognized A's intention differently, B would not make the response that B makes to A's utterance. In this sense, recognizing A's intention is why B does what B does. Lewis's account of convention thus dovetails with Grice's analysis of meaning.

Despite these attractive features of Lewis's appeal to convention, the understanding of language that emerges from it is problematic in many respects. In the next few subsections, we will explore some of these problems.

20.3.1 Improvised Meaning

Lewis is working, at least implicitly, with a view of language as relatively stable and in common currency. On this view, semantics reveals widespread regularities that hold across people, space and time. The meanings of words might be a prototypical example. Because such regularities are widespread and stable, there won't be anything tendentious about thinking about language use in terms of recurring coordination problems and conventional solutions to them.

This way of thinking about language is harder to maintain if improvisation is an essential part of meaning in language. When we coin new words, modulate established meanings, and arrive at creative understandings of one another, we do not rely entirely on conventional semantic properties. Indeed, we may not even take up any expectations that we are going to continue to use words in these new ways. If we want to characterize meanings in such cases, it seems unlikely that we can do so just with shared solutions to recurrent coordination problems.

Indeed, even for Lewis, conventions are established gradually. At first, coordination succeeds through other mechanisms, like salience, good luck, or partial or tentative precedents. In these cases, speakers don't yet have common knowledge of an equilibrium that would settle the meanings of their words. We cannot appeal to convention to characterize what their meanings are in these cases.

Indeed, the more you think about it, the more problematic it seems to try to *reduce* knowledge of meaning to knowledge of other agents' strategies for using language, as Lewis seems inclined to do. This becomes particularly clear if we consider the improvised meaning that Grice was most concerned with, so-called pragmatic "conversational implicatures" (Grice 1989). We can separate semantics from other aspects of utterance interpretation only if we can distinguish what an utterance means from everything that the audience can infer from the fact that the speaker has said it. But Lewis is out to explain meaning in terms of our expectations about one another. For Lewis,

utterances get their meanings from the prevailing mutual expectations about what everyone is going to do. In the signaling games which capture the essence of Lewis's idea, what an agent learns when she receives a signal is exactly that the world is the type of place and the speaker is the kind of person who would send that signal now. That's never going to distinguish semantics from pragmatics in an intuitive way.

Take a concrete case. We want to assign different meanings to "There's a bug on your back!" and "Swat the bug!" But each would be uttered when there's a bug on the hearer's back, and each calls for the hearer to swat that bug. So, each has an equal claim to mean, there's a bug, and each has an equal claim to mean, swat it!

Lewis recognizes this problem, and has a particular project for how to deal with it. He thinks that his conventions of truthfulness and trust in natural language aren't subject to exactly these criticisms. The conventions for meaning in natural language, he argues, are productive and systematic in ways that resolve much of the indeterminacy we see in the Paul Revere case. In other words, "There's a bug on your back!" and "Swat the bug!" may be used in the same circumstances, but we will ultimately have to give them different conventional meanings, because we will ultimately have to have a compositional semantics, and compositionality will distinguish them.

Compositionality certainly places profound constraints on semantics. But if much of meaning is in fact improvised, then we will always need to describe institutions that let us make meaning, not just institutions that let us use established meanings. Even if we keep Lewis's central place for coordination and social competence in delimiting semantics, we will need some additional tools. And this broader account is likely to fundamentally shape the way we see the boundary between semantics and pragmatics. We can describe the difference between "There's a bug on your back!" and "Swat the bug!" *not* in terms of general expectations of when speakers use utterances and how their audiences react, but in terms of the specific dynamics and institutions of meaning that make one an utterance about how things are and the other an utterance about what the audience should do.

20.3.2 Universal Grammar

Lewis is also working, at least implicitly, with a view of language as entirely learned. This shows up in the very strong sense in which conventions are arbitrary in Lewis's account. When we have a convention to solve a coordination problem with a certain strategy, we've learned that people around us follow that strategy – or we've worked to establish it as an equilibrium with the people around us. Moreover, if we have a convention, then we could have arrived at a different joint strategy had our experiences and choices been different. This also shows up in the strong assumptions that Lewis makes about the mutual knowledge and rationality associated with linguistic conventions. Because Lewis views linguistic conventions, at least implicitly, as learned manifestations of people's general social competence, Lewis says that a convention is established only when agents have common knowledge of one another's preferences, one another's rationality, one another's likely choices, and hence the convention itself. These are very strong assumptions, indeed, but they are necessary to explain why the convention gives agents general and decisive reasons to choose the actions they do.

We've seen that this is not the only reason why one might observe a particular equilibrium in a coordination game. In particular, in some circumstances, we expect equilibria to arise through *arbitrary psychological mechanisms*, rather than through arbitrary experience. Linguists in the Chomskyan tradition take precisely this sort of approach to linguistic communication. On this view, significant aspects of linguistic structure and meaning are universal across human languages. They are determined by the language faculty, the underlying cognitive module that enables people to learn and use their native language. The operation of the language faculty helps to explain how children are able

to acquire their native language so quickly, simply by participating in the highly limited linguistic interactions of early life. What children need to learn is just relatively superficial particulars about the language spoken in their community. See Pesetsky (1999) for a brief overview or Baker (2002) for a more detailed exposition of this perspective.

Linguists' understanding of the language faculty fits poorly with the assumption that linguistic meaning is conventional in Lewis's sense. Speakers' linguistic choices are not always the product of their deliberative rationality; they are often automatic consequences of the operation of language faculty.

If speakers' choices match, it may be because the same mechanisms apply across individuals, not because speakers make their choices by taking learned expectations about one another into account. In this case, speakers could act in accordance with the rules of language without even *knowing* what those rules are – much less having common knowledge of the rules. In particular, we might question whether speakers need have any beliefs, even tacitly, about how their own language faculty operates; the language faculty simply does its thing. But suppose we do acknowledge an attitude that speakers bear to the information that tacitly characterizes the language faculty. If that information simply reflects the innate architecture of the system, we have little reason to characterize speakers' attitudes epistemically, in terms of justified conclusions informed by relevant evidence. See for example Chomsky (1980) or Devitt (2006).

Note, however, that this understanding of the language faculty may still make crucial appeal to the more general notion of social competence involved in successful coordination. Children must acquire the specific facts that distinguish their native language from other possible languages. These facts are not determined just by the practical interests of members of the community, but crucially depend on the alignment that underwrites coordination. This fundamental arbitrariness constrains the evidence that children should use and the hypotheses that they should entertain in coming to understand the language spoken around them. The reasoning involved may well be different from what children need to understand actions that accomplish practical goals by more straightforward means.

The Chomskyan view provides an interesting foil to Quine's objections about language and convention. If we say, with Chomsky, that some rules of language are a manifestation of an innate social competence, there can be no fear of regress in how these rules arise. Our innate social competence does not have to be agreed on in words; indeed, there is no way it could be. Our innate social competence simply evolves. As new features are introduced into the underlying architecture, they trigger corresponding universals in the linguistic capacities that the architecture manifests.

We needn't go as far as the Chomskyan picture to worry about applying a notion of convention as strong as Lewis's to linguistic knowledge. Burge (1975) offers a simple case. Consider an isolated population, whose members mistakenly subscribe to the view that their language is the only possible language. This no longer satisfies Lewis's definition of a convention. To satisfy that definition, the group has to have mutual knowledge that there's another alternative equilibrium that they could have adopted. This is exactly what Burge's isolated population denies.

This discussion again suggests Lewis is asking too much in reducing meaning to convention. His notion of coordination provides a useful handle on what's distinctively social about language, particularly meaning in language. When agents face coordination problems, they can't succeed just by practical rationality, because practical rationality doesn't dictate which of the many equilibrium strategies is the right one. So, agents that do coordinate successfully must have additional processes at play – innate, heuristic, or deliberative – that allow them to align their choices to one another. Speakers of language do coordinate. Explaining meaning in language thus means getting clear on the processes that the agents must have – innate, heuristic, or deliberative – that lets them match their signals and their interpretations.

Conventions, as Lewis considers them, are one possible explanation. But it seems likely that there are crucial ingredients of meaning in language that can't be captured this way. The problem cross-cuts the problem of improvised meaning that we considered earlier. Improvised meanings are not conventional because they are not yet established; innate meanings are not conventional because they are not established through the right mechanisms. We draw a similar moral about Lewis's approach from the two cases, however. We believe (and argue as much in Lepore and Stone forthcoming) that once we say more about the kind of coordination involved in semantics, we will feel less pressure to go further still, and demand that this coordination necessarily be based in convention.

20.3.3 Other Perspectives on Convention and Meaning

Lewis's discussion of coordination and convention has sparked a wide- ranging philosophical inquiry into how far rationality, experience, and mutual expectations actually go in explaining what people do together. For example, philosophers have pointed out social customs of various kinds that do not fit Lewis's notion of convention. People often seem to stick to customary patterns of behavior, but more out of habit than out of preferences that actively require conformity (Gilbert 1989; Millikan 1998). At home, the family table normally accommodates a few guests. Still, when the family sits down to eat, everyone may routinely pick the same customary place to sit and leave the same seats empty. It's open to everyone to mix things up, and it might be interesting or even preferable to do so from time to time. So, when everyone persists in doing what they usually do, it's not a convention in Lewis's sense. It's possible that uses of language could likewise be routine but not conventional, in theoretically significant ways. For example, you might think this way about sticking to the familiar formulation in indirect speech acts like "Can you pass the salt?"

Philosophers have also argued that some customs continue to figure in the cultural repertoire despite the fact that people act in accordance with them only rarely. An example is handing out cigars at the birth of a baby. This is not, let us suppose, what most new parents actually do nowadays. So, there are no prevailing mutual expectations that new parents will celebrate by handing out cigars. And parents' choices to hand out cigars, if they do so, cannot respond to the expectations of the community that they will. Handing out cigars is not a convention, in Lewis's sense. But it is still a customary option, whose cultural meaning is shaped by the weight of precedent. We might want to accommodate such options, and explain their relationship to other kinds of customary behavior, in a broader theory of the social world. The category might prove just as useful in accounting for elements of language, such as rarely used verbal clichés.

A different sort of worry has to do with Lewis's requirement that the interests of parties to conventions coincide. There seem to be exceptions, including many of the cases Hume was originally interested in, such as accounting for the institution of property as a convention (Vanderschraaf 1998a). Considering more general situations exposes two weaknesses in Lewis's account of conventions of meaning. The first has to do with Lewis's characterization of the mutual expectations involved in convention. In general games, as in Lewis's coordination games, there can be multiple alternative Nash equilibria. Again, in general games, when an agent chooses the strategy involved in one Nash equilibrium R rather than another R′, the agent's expectations about what others will do are decisive. The agent prefers to participate in R rather than R′ given that others do. However, in general games, it may no longer be the case that the agent playing R prefers that others play R as well. Perhaps, the agent would be better off if others capitulated and simply abandoned their own interests. (Of course, that won't happen.) A more general account of convention must therefore adjust Lewis's definition not only to drop the requirement that the interests of the parties to the convention coincide, but also to characterize agents' mutual expectations in appropriate game-theoretic terms. See Sugden (2004) or Vanderschraaf (1998b).

More specific problems arise with signaling in the presence of conflict. If signals are cheap to produce and the interests of the sender may differ from those of the recipient, then the equilibrium strategy of the receiver is often not to trust the signal, but simply to ignore it. The recipient may have good reason to expect that she would get the same message from the sender no matter what the world was like. In one-shot games, adversarial communication seems to depend on the use of costly signals, which carry penalties that outweigh the benefits deceptive senders could possibly get from using them. Recipients can trust costly signals because no sender has an interest in using them unreliably. Seminal work on costly signals includes Spence (1973) in economics and Grafen (1990) in biology.

Utterances do not seem to be costly signals, so we would require a different tack to link meaningfulness to the strategies of language users with conflicting interests. One way to salvage meaningful cheap signals is to consider the interactions of senders and receivers who must talk and act repeatedly over time. When agents interact indefinitely with one another, a wide range of equilibria is possible. This is known as the *general feasibility theorem* or *folk theorem* in economics. These equilibria work because repeated games allow for the threat of punishment. Agents stick to the expected course of action because they know that, if they deviate in any one round to pursue their short-term self-interest, other agents will retaliate in subsequent rounds and impose long-term penalties that outweigh their short-term gains.

Suppose that conflicting interests in society make this the right way to think about communication. Then not only would we need to adjust Lewis's definition of convention, perhaps in line with Sugden (2004) or Vanderschraaf (1998b), we'd also need to adjust our understanding of what the conventions that underwrite meaning actually amount to. They'd be not just conventions of truthfulness and trust, following Lewis, but conventions of truthfulness, trust, and punishing liars.

People certainly do retaliate against liars. But lies are particularly problematic only if, like Lewis, we attempt to characterize meaning exclusively through interlocutors' expectations about when speakers will use utterances and how addressees will respond to them – that is, directly in psychological terms. If we think more abstractly in terms of activities of making ideas public or reaching agreements about how things are, we don't put ourselves in the position of having to account for the content of a lie in terms of the circumstances in which a liar uses it, or of explaining the joint interest that interlocutors have in a productive conversation in terms of interests they share in practical outcomes. So, while we acknowledge that conflict has an important place in a broader account of conventions and social competence, we think that this alone does not prevent us from linking meaning to coordination, following Lewis.

Finally, social competence seems to play a role not just in explaining the strategies that we follow when we undertake joint activities but in explaining how many of those activities are constituted in the first place. Think of a game like chess. Players have to coordinate when they play chess. The satisfaction they derive from the game comes in part from the shared standards they must apply to interpret the moves and track the flow of play. Moreover, these rules seem arbitrary in many respects. We could agree to change the rules in a wide variety of ways. We would get different games, perhaps all quite entertaining, with varying degrees of similarity to official chess. In this sense, when players agree to play a game of chess, they implicitly adopt the convention of interpreting their moves according to the rules of chess. So, Lewis's account of social competence is an important part of explaining how chess works.

However, the example of chess also shows the limits of Lewis's model of the origins of conventions (Marmor 2009). Lewis's model seems suited to cases like driving where agents face a coordination problem that exists antecedently and must converge on a convention for solving the problem. They muddle through, eventually relying on the weight of salience and precedent to secure reliable mutual expectations about what to do.

Imagine trying this with chess. Players start playing a game. They improvise their moves and their meanings, sometimes succeeding in arriving at a common understanding of the state of play through the weight of salience and precedent. Eventually, they have a shared understanding that they are playing chess. Explaining the conventions this way seems to miss the point: the problem of playing chess is itself the product of agents' interaction. Players need to converge on the rules of the game, not just on their strategies for playing. They can do this only if they have the right antecedent relationships and institutions, so that they can flesh out, refine and litigate their inchoate rules to cover fairly the cases that arise as they play. Similarly, to the extent that interlocutors are engaged not just in using existing meanings but in establishing new meanings, they too will require not only conventions governing existing meanings but also the appropriate relationships and institutions for pursuing new meanings.

We have already noted, however, that meaning can be improvised. Improvised meaning is coordinated, but not conventional. Universals of grammar are not conventions either. Focusing on the social competence needed for communication seems like a better way to update Lewis's insights than to respond to a litany of putative counterexamples to Lewis's account of conventions in general and conventions of meaning in particular. Lewis's insight linking coordination and communication remains central to this project. But the dialectic is much changed from Lewis's original one. Gone is the goal of characterizing meaning exclusively in terms of agents' expectations about one another's strategies. Gone is the worry that such expectations might be impossible to establish without a language to express them in. Instead, the question is how we can characterize the enterprise of communication, to naturally highlight the coordinated strategies interlocutors are following in producing and understanding utterances, and the diverse linguistic resources that are implicated in these strategies.

20.4 Conclusion

It's hard to imagine in retrospect, but before Lewis' work in *Convention* the notion of a convention was deeply mysterious. Here is another debt the profession owes Lewis: thanks to his concepts and analytical tools, we not only understand how conventions can be the natural outgrowth of our interactions with one another, but we can in fact pursue sophisticated inquiry into the nature and scope of those conventions. Lewis's contribution is achieved through ideas whose influence has only increased over time – his account of coordination remains crucial to our understanding of strategic interaction, signaling games continue to provide a fundamental tool for understanding communication, and theorists still must acknowledge that the achievement of coordinated outcomes by rational choice depends on the common knowledge that Lewis associates with convention. In this sense, Lewis's contribution stands as a testament of philosophy not only to find the right questions but to answer them as well.

References

Baker, M. (2002). *The atoms of language*. New York: Basic Books.
Burge, T. (1975). On knowledge and convention. *Philosophical Review*, 84, 249–55.
Chomsky, N. (1980). *Rules and representations*. New York: Columbia University Press.
Clark, R. (2012). *Meaningful games: Exploring language with game theory*. Cambridge, MA: MIT Press.
Devitt, M. (2006). *Ignorance of language*. Oxford: Oxford University Press.
Gilbert, M. (1989). *On social facts*. London: Routledge.

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology*, 144(4), 517–46.
Grice, H.P. (1957). Meaning. *Philosophical Review*, 66(3), 377–88.
Grice, H.P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.
Hume, D. (1740). *A treatise of human nature: Being an attempt to introduce the experimental method of reasoning into moral subjects*. London: John Noon, http://www.gutenberg.org/ebooks/4705, accessed June 3, 2014.
Lachmann, M., Számadó, S. & Bergstrom, C.T. (2001). Cost and conflict in animal signals and human language. *Proceedings of the National Academy of Sciences of the United States of America*, 98(23), 13189–94, http://www.pnas.org/content/98/23/13189.full, accessed June 3, 2014.
Lepore, E. & Stone, M. (forthcoming). *Imagination and convention*. Oxford: Oxford University Press.
Lewis, D.K. (1969). *Convention: a philosophical study*. Cambridge, MA: Harvard University Press.
Marmor, A. (2009). *Social conventions: from language to law*. Princeton, NJ: Princeton University Press.
Maynard Smith, J. & Price, G.R. (1973). The logic of animal conflict. *Nature*, 246, 15–18.
Millikan, R.G. (1998). Language conventions made simple. *Journal of Philosophy* 95(4), 161–80.
Nash Jr., J.F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences of the United States of America*, 36(1), 48–9, http://www.jstor.org/pss/88031, accessed June 3, 2014.
Pesetsky, D. (1999). Linguistic universals and universal grammar. In *The MIT encyclopedia of cognitive science*, pp. 476–78. Cambridge, MA: MIT Press.
Rescorla, M. (2011). Convention. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2011 edition), http://plato.stanford.edu/archives/spr2011/entries/convention/, accessed June 3, 2014.
Russell, B. (1921). *The analysis of mind*. London: Allen and Unwin.
Schelling, T.C. (1960). *The strategy of conflict*. Cambridge, MA: Harvard University Press.
Skyrms, B. (2010). *Signals: evolution, learning and information*. Oxford: Oxford University Press.
Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3), 355–74.
Su, S., Cai, F., Si, A., Zhang, S., Tautz, J. & Chen, S. (2008). East learns from west: Asiatic honeybees can understand dance language of European honeybees. *PLOS ONE*, June 4, http://dx.doi.org/10.1371%2Fjournal.pone.0002365, accessed June 3, 2014.
Sugden, R. (2004). *The economics of rights, co-operation and welfare*, 2nd ed. New York: Palgrave.
Vanderschraaf, P. (1998a). The informal game theory in Hume's account of convention. *Economics and Philosophy*, 14, 215–47.
Vanderschraaf, P. (1998b). Knowledge, equilibrium and convention. *Erkenntnis*, 49, 337–69.
von Frisch, K. (1967). *The dance language and orientation of bees*. Cambridge, MA: Harvard University Press.
von Neumann, J. & Morgenstern, O. (1947). *Theory of games and economic behavior*, 2nd ed. Princeton, NJ: Princeton University Press.

21

Asking What a Meaning *Does*

David Lewis's Contributions to Semantics

BARBARA H. PARTEE

21.1 David Lewis's Background and Early Involvement with Linguistics

David K. Lewis was a central player in the golden age of linguistics–philosophy interaction, and he and his work were important parts of what made that age golden. According to Brian Weatherson's article on David Lewis in the *Stanford Encyclopedia of Philosophy* (Weatherson 2010), his greatest contributions were to metaphysics and to the philosophy of mind, with additional significant contributions to a whole list of fields that includes the philosophy of language. That may be so; but his contributions to the philosophy of language and to semantics are enormous and influential, and, of course, for linguists, those contributions are what we know best and cite most, and those are what I am writing about in this chapter. But first some background, including some shared personal history.

David did his undergraduate degree at Swarthmore, where we both started in 1957, and where I first got to know him a little bit. He spent the 1959–60 year at Oxford, where he had Iris Murdoch as his tutor and attended lectures by Ryle, Grice, Strawson, and Austin, and where he reportedly fell in love with philosophy. When he returned to finish his last two years at Swarthmore, he decided to major in philosophy, not chemistry as he had originally intended. I think we were together in just one or two philosophy seminars at Swarthmore – enough for me to be mightily impressed, and to appreciate that he wasn't just that shy guy who shuffled through the hallways without making eye contact with anyone.[1]

At Swarthmore there was no linguistics back then. But during the summer of 1960 there was a summer seminar in "Structural Linguistics" taught at the University of Pennsylvania by the Polish linguist and logician Henry Hiż for students with backgrounds in mathematics, philosophy, or psychology. Three of us from Swarthmore found our way into it: David, Gil Harman, and me. Jerry Katz and Jerry Fodor were there with their freshly minted PhDs in philosophy from Princeton, and they regularly sat in the back and kibitzed. What we learned about and worked with in that seminar was the version of transformational grammar developed by Zellig Harris, Chomsky's PhD advisor.[2] As far as I know, that was David's first introduction to linguistics, as it was mine.

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

David graduated from Swarthmore in 1962 and went to Harvard in philosophy; I graduated in 1961 (in mathematics) and went to MIT in linguistics. Our paths continued to cross in Cambridge – David sometimes came to Chomsky's lectures, and I took one course from Quine, and we got to know each other better during those years than we had at Swarthmore.

We both went to UCLA for our first jobs, coincidentally. I joined the Linguistics Department there in 1965, and David joined the Philosophy Department in 1966. As Stephanie Lewis recalls,[3] UCLA was the best combination for David as a beginning assistant professor and for her as a beginning philosophy graduate student. I started out as very much a "Chomsky disciple" and a syntactician, but like several others of my generation, I was increasingly interested in semantics and was gradually discovering how much ground logicians and philosophers had already covered. Linguists in the late 1960s were engaged in the Generative Semantics/Interpretive Semantics wars (Partee 1970; Newmeyer 1980; Harris 1993). I know from conversation and correspondence that David was quite sympathetic to much of what the Generative Semanticists were doing, especially George Lakoff and James McCawley; that can also be seen in some of his remarks in "General Semantics" (Lewis 1970), which I'll say more about in section 20.2.

David stayed at UCLA for a relatively short time,[4] 1966–70, but his UCLA period was an important one for the field of semantics. His colleagues during those years included Richard Montague, David Kaplan, Keith Donnellan, Rudolf Carnap, and Alonzo Church. Hans Kamp was there at the beginning, finishing his PhD under Montague, although I didn't meet him until later. And the famous UCLA "logic year" was in 1967–8, David's second year there, with visitors including Wilfrid Hodges, Jon Barwise, and Jerome Keisler. Max Cresswell, who became a close friend and colleague, visited in the winter quarter of 1970, hosted by David and Steffi (both of whom he was meeting in person for the first time) until he found a place to live. It was at UCLA that David rewrote his dissertation into his book *Convention* (Lewis 1969); I recall him checking with me on a phonetic transcription in one of the last two chapters. (I had to check with a real phonetician to be sure; David had it exactly right.)

And it was sometime in 1968 that David first put me onto Montague's work. I had seen Montague in some Philosophy Department colloquia, and knew of his sharp tongue but knew nothing of his work. David told me that Montague was starting to apply some of his work in logic and recursive function theory to the semantics of natural language. David knew me well enough to guess that I would find it interesting to sit in on Montague's seminars, which I did for the first time in 1968, along with him and my PhD student Frank Heny. In post-seminar discussions, David helped us greatly in deciphering the logic and appreciating the ideas.

David had discussions with Montague, and his own ideas in "General Semantics" share some properties with Montague's work; they seem to have influenced one another a bit, although they were never collaborators and their main interests and motivations were quite different. David regretted that Montague did not hold a very high opinion of linguists and did not consider it worthwhile to pay much attention to what Chomsky or other generative grammarians were doing. Happily, David had a much more congenial and mutually appreciative relationship with linguists.

While it was Montague's work that formed the principal starting point for the development of formal semantics, David did much more than Montague to actually bring linguists and philosophers, and their ideas, together. In a letter to me in November 1969, while I was away from UCLA on sabbatical at MIT, David wrote,

> Where will I find a precise definition of a transformational grammar? Peters and Ritchie 1969 [eventually published as Peters and Ritchie 1973], presumably, but I don't know where or what that is. Montague's been reading *Aspects* [Chomsky 1965] and Jacobs & Rosenbaum [Jacobs and Rosenbaum 1968], and claims he still hasn't the foggiest idea what a transformation is; I'd like to try to shut him up, though it's probably not a worthwhile thing for me to try to do.

He continues that Montague was planning to read a new version of "English as a Formal Language" [Montague 1970b] at the upcoming APA-ASL meeting in New York. "Having read a little transformational grammar, he plans to lambaste it more than ever. I'd like to [sic] Postal or someone like that on him, but I don't know Postal; but maybe you'd like to publicize the event among the more contentious types at MIT." And in a letter written in 1972, some time after Montague's death, he refers to "Montague's efforts to isolate his approach from others' thought about the same problems."[5] There were others who also tried to persuade Montague to pay more attention to Chomsky's work, notably Yehoshua Bar-Hillel and Frits Staal,[6] but it was only after Montague's death that the two traditions were brought into real cooperation, through the efforts of David, Rich Thomason, Terry Parsons, the present author, and others.

David's earliest work directly relevant to linguistics was his dissertation and the published book *Convention* (Lewis 1969). But what really "reached" linguists first, and spoke directly to them, was his "General Semantics," which began as a paper presented at the Third Annual La Jolla Conference on Linguistic Theory and the Structure of English in March 1969. Those present included leading Generative Semanticists George and Robin Lakoff, James McCawley, Haj Ross, Paul Postal, leading Interpretive Semanticists Ray Jackendoff and Joe Emonds, as well as Emmon Bach, Ed Klima, Yuki Kuroda, Ron Langacker, Leonard Newmark, Barbara Partee, Robert Stockwell, and Paul Schachter. The paper was widely circulated and discussed in pre-publication versions before its actual publication date of 1970. For many of those in attendance, this was their first encounter with David Lewis, and a memorable one.

From that time onward, David was often invited to conferences that included both philosophers and linguists, and was as highly valued for his contributions to the discussion of papers as for his own papers. I will return in the final section of the paper to the topic of David's important role in helping linguists and philosophers understand each other and cooperate in working on the syntax and semantics of natural languages.

In the rest of this chapter, I will discuss some of his ideas that have been most important for linguists, especially but not only for formal semanticists, and some of the influence his ideas have had. The papers that have been most influential among linguists and are most cited are "General Semantics" (Lewis 1970), "Adverbs of Quantification" (Lewis 1975b), "Scorekeeping in a Language Game" (Lewis 1979b), "Attitudes *De Dicto* and *De Se*" (Lewis 1979a). Especially important for foundational issues in semantics were his book *Convention* (Lewis 1969) and the related paper "Languages and Language" (Lewis 1975a), which are discussed in Ernie Lepore's and Matthew Stone's contribution to this volume (chapter 20). The "Scorekeeping" paper is central to Craige Roberts's contribution to this volume (chapter 22). Other papers of Lewis's that have been important for particular topics in semantics are "'Tensions" (Lewis 1974), "Probabilities of Conditionals and Conditional Probabilities" (Lewis 1976) (and his other work on conditionals), "Truth in Fiction" (Lewis 1978), which he was working on during the 1974 Linguistic Institute in Amherst, and "Index, Context, and Content" (Lewis 1980), his Uppsala 500th centennial conference paper. My discussion will be organized primarily around ideas and secondarily around the papers where they were expressed, starting in section 21.2 with the idea from which I took the title of this chapter, which appears very near the beginning of his first most influential paper for linguists, "General Semantics." Section 21.3 takes up Lewis's distinction between a language as an abstract object and the use of a language by a community, and its importance for defending the non-psychologistic foundations of formal semantics. The brief section 21.4 mentions some issues of importance to linguists connected with Lewis's work on possible worlds, modal realism, conditionals, and counterpart theory, brief because I leave the heart of that work to discussion in other contributions. Section 21.5 highlights a number of specific ideas that have influenced linguists, and the concluding Section 21.6 adds some general remarks about David Lewis's role in helping linguists and philosophers to understand and appreciate one another.

What can't be captured in a report on his ideas and their influence is the way every conversation with David Lewis was a memorable experience. If you asked David a question, there might be a silence while he thought. And then he would give his thought-out reply in complete sentences organized into good paragraphs. And it wasn't as if he were giving canned or rehearsed responses. When he answered my questions, he always understood my background and what had motivated my questions; he was giving an explanation for me. I loved the years when we were both sitting in on Montague's seminars and I could ask him questions afterwards, and I probably learned as much from David as from Montague. And, in later years when we would be together at conferences, his questions and comments were often among the highlights of the conference for me. I am sure I am not alone in such experiences.

21.2 Lewis's Central Methodological Advice for the Study of Meaning

The idea of David Lewis's that I have quoted most often in my writing and in my teaching is this one, from page 22 of the original *Synthèse* publication of "General Semantics" (1970) (henceforth GS), which I generally refer to as "Lewis's advice":

(1) Lewis's advice: In order to say what a meaning is, we may first ask what a meaning does, and then find something that does that.

That advice really helped to revolutionize the field; it forced linguists to become more conscious of their unexamined ideas about meaning, and it offered promise of making meaning something that could be as tractable and as "scientific" as syntax. It also helped to soften up linguists to the initially very foreign idea of identifying meanings, or at least a central part of meanings, with truth conditions. That strategy has remained influential, and one can see it at work implicitly or explicitly in many subsequent proposals for enriching or replacing the truth-conditional conception of meaning. The proposals by Kamp and Heim in the 1980s for a more dynamic conception of meaning, for instance, were argued on the basis of some things that truth-conditional theories had been unable to solve effectively, including the semantics of donkey sentences and the behavior of anaphora in discourse.

Lewis in GS proceeded to make a range of proposals and then to address many questions that might be raised concerning various aspects of what he had just proposed, showing how they might be resolved. The calm and methodical manner in which he went about that also had an influence on all subsequent work. In principle, although unfortunately not in all people's practice, he removed arguments about what meanings are from the realm of dogma. That reflects how non-dogmatic David was in general; in many papers he included parts about different ways that you might do things, even while indicating and arguing for the way he thought was best. There are repeated examples of that in GS, also in Lewis (1974, 1980), and in numerous other works.

One of the first applications of Lewis's advice in GS is the defense of truth conditions as central for the study of meaning, and the rejection of Fodor, Katz, and Postal's "Markerese" (Katz and Fodor 1963; Katz and Postal 1964). Another famous quotation from the same paper:

(2) But we can know the Markerese translation of an English sentence without knowing the first thing about the meaning of the English sentence: namely the conditions under which it would be true. Semantics with no treatment of truth conditions is not semantics. (1970, 18)

It must be noted, of course, that the principle is used by many *against* formal semantics, arguing that treating truth conditions is *not* the main thing we want meanings to "do" (Lakoff 1987; Jackendoff 1996). There are two common responses. One is an ecumenical one: anyone is welcome to declare their own ideas about what they want meanings to do and pursue an account of meanings that will do that. Another is a "strategic" response: of course there is probably much more to meaning in a broad sense than truth conditions, but truth conditions must be part of any full account for all those sentences that have truth conditions, and getting those right is a challenging but relatively well-defined task, and therefore a good thing to work on.

The methodological point encapsulated in Lewis's advice shines through much of Lewis's work on language. It was used to argue for the need to distinguish index from context (Lewis 1980), as well as to discuss what needs to go into contexts (Lewis 1979b).

One important and influential application of that same principle, already in GS, was to point to the need for something richer than intensions as the meanings of embedded sentences in propositional attitude contexts. On page 25 of GS, he writes:

> **(3)** Intensions, our functions from indices to extensions, are designed to do part of what meanings do. Yet they are not meanings; for there are differences in meaning unaccompanied by differences in intension. It would be absurd to say that all tautologies have the same meaning, but they have the same intension . . . Intensions are part of the way to meanings, however, and they are of interest in their own right. We shall consider later what must be added to an intension to obtain something that can do all of what a meaning does.

He then introduces the idea of structured meanings, without that name (he just calls them "meanings"), with reference to Carnap (1947) on "intensional isomorphism" and to C.I. Lewis (1944) on "analytic meaning."

> **(4)** It is natural, therefore, to identify meanings with semantically interpreted phrase markers minus their terminal nodes: finite ordered trees having at each node a category and an appropriate intension. (1970, 31)

This includes for any compound expression the categories and intensions of its parts, and their parts, etc. But then he says this might cut meanings too finely; it would leave us "unable to agree with someone who says that a double negation has the same meaning as the corresponding affirmative." He says not to worry; we can use intensions for some purposes and meanings for others, and perhaps something of intermediate fineness can also be found. This has remained an issue, of course. But the idea that some such kind of structured meanings may be part of the solution to the problem of the semantics of propositional attitude constructions, and useful in other ways as well, was followed up on in a number of ways (von Stechow 1984; Cresswell 1985; Chierchia 1989a).

In section 21.5 I will say something about the syntactic architecture proposed in GS, a transformational grammar with a categorial base component. Here I want to note that, as another example of how he often provided alternatives to his preferred ideas, he shows how with the conception of meanings just introduced, grammars can be reconstrued as conditions on well-formedness for meanings (citing McCawley 1968), where it was shown how phrase structure grammars could be reconstrued as tree-checkers), with meanings construed as trees. He also showed how many other notions could be defined in terms of his structured meanings, one fundamental notion being the truth of a sentence meaning at an index.

In various places in GS he takes pains to point out choices that are left open by his proposals. On page 31, he notes that all that he has said about intensions for derived categories and about the form

of projection (composition) rules "is independent of my account of intensions for basic categories. Whatever S-intensions and N-intensions may be – even expressions in Markerese or ideas in someone's mind – it is still possible to take S/N intensions as functions from N-intensions to S-intensions and to obtain the intension of "Porky grunts" by applying the intension of "grunts" as function to the intension of "Porky" as argument." This is important – he was arguing that one can accept the goal of building a compositional semantics with the help of function-argument structure whether or not one accepts his objections to semantics without truth-conditions. (The same was true for Montague in his "Universal Grammar" (Montague 1970a), where he argued that syntax must be an algebra and semantics must be an algebra and there must be a homomorphism from the syntactic algebra to the semantic algebra, but left it very much open what the nature of the elements of those algebras might be.)

He also noted, surprisingly to some of us, that his proposals were neutral with respect to the "linguistic wars" then going on in full force between Generative Semantics and Interpretive Semantics.

> **(5)** I have said nothing to eliminate generative semantics. What I have chosen to call the "lexicon" is the initial lexicon. . . . In might be that none of the initial lexical items ever reach the surface . . . In that case it would be appropriate to rechristen my base structures "semantic representations". In that case also there might or might not be a level between base and surface at which word-introducing transformations are done and other transformations have not yet begun.
>
> I have also said nothing to eliminate surface semantics. This may seem strange . . . However, I rely here on the observation (Lakoff 1971) that surface-structure interpretation rules are indistinguishable from global derivational constraints relating three levels . . . And we might have constraints relating base structure not only to "deep" and surface structure but also to structure at various other intermediate levels. (1970, 38)

In another part of the paper, in response to frequent objections that the meanings of non-declarative sentences cannot be treated in terms of truth conditions, he shows how an approach like that of Stenius (1967) could work, letting a meaning for a sentence be "a *pair* of a mood and an S-meaning (moods being identified with some arbitrarily chosen entities)" (1970, 56). On a Stenius-style approach, there would be a new root node SENTENCE that would have as constituents "MOOD" and the earlier root node "S"; Ss are then "sentence radicals", not declarative sentences, but all the old semantics of truth-conditions applies to them. The semantics of mood consists, according to Stenius, of rules of language *use*. Lewis's response to Stenius's *use*-semantics is to keep what he refers to as "abstract semantics" distinct from a theory of language use, and therefore to include "some arbitrarily chosen entities" as place-holders for moods.

Lewis himself in GS preferred an alternative method, which seems to be inspired by work in Generative Semantics, especially (Ross 1970).[7] His idea is to make use of certain distinguished classes of base structures of roughly the form of performatives (*I ask you whether S*, etc.), and then to treat the *non*-declaratives as paraphrases of the corresponding performatives. But, unlike Ross, Lewis didn't want to take an ordinary declarative sentence S as a paraphrase of a performative sentence *I declare to you that S*, since he considered the two to have different truth-conditions.

I close this section with two recent quotations advocating different ways of doing semantics based on different ideas of what a meaning (should) do – they reach conclusions quite different from Lewis's, but I think they approximately follow the idea that you figure out what you think a meaning should do and then build a theory around something that does that.

(6) A quotation from the introduction to van Lambalgen and Hamm (2005):

The formal setup we have chosen reflects our belief that semantics, in order to be cognitively relevant, should start from a computational notion of meaning. Using the traditional terminology of sense and denotation, our point of view can be described succinctly thus: "the sense of a linguistic expression is the algorithm which allows one to compute the denotation of that expression."

(7) A quotation from Pietroski (2003):

If this is correct – if our best semantic theories turn out to be theories of linguistic features that do not determine truth-conditions – and meaning is what our best semantic theories are theories of, then the meaning of a sentence doesn't determine its truth-condition. Sentences, as products of (largely innate and modular) language systems, have truth-conditions only by virtue of their relation to other cognitive systems and the environments in which the sentences are used. But sentences have their meanings by virtue of more local facts concerning the psychology (and hence biology) of language-users. So a semantics that makes the right theoretical cuts will not itself associate sentences with truth-conditions.

21.3 Languages and Language

Another point of Lewis's that I have always found important and insightful is the importance of distinguishing between a language as formally describable object and the "psychological and sociological facts" that make a language someone's language. That point is given close attention in *Convention* (Lewis 1969) and in "Languages and Language" (Lewis 1975a), and is discussed briefly in GS.

This distinction is a foundational one; it applies not to particular semantic analyses but rather addresses and replies to those who advocate a view of meaning as use or of meaning as understanding. Formal semantics does not offer a theory of use, and Lewis argues that that is as it should be. Nor is it a theory of understanding an utterance, nor more abstractly, of "semantic competence."

I worried in some earlier papers (Partee 1977, 1979) about how one might reconcile the Fregean non-psychologistic view of semantics standard in formal semantics with the Chomskyan view of linguistics as a branch of psychology, an account of the unconscious linguistic competence of a native speaker of a language. It was Lewis's work that most helped me understand that one might consider Chomsky's syntactic theory without his psychologism, and separate the technical challenge of combining Montague's or Lewis's semantics with Chomsky's syntax from the "psychological" question of how to characterize knowledge of a language (syntax and semantics) and its acquisition. Lewis's views on these issues are reflected in "Intensions and Psychological Reality" (Dowty 1979, chapter 8); Dowty confesses that he believes "that the model-theoretic intension of a word has in principle *nothing whatsoever* to do with what goes on in a person's head when he uses that word." Dowty appeals to arguments from Cresswell (1973), from "Reference and Understanding" (Putnam 1978, part 3), and from Lewis (1969). And those ideas of Lewis's were central for me when I commented (Partee 1988) on Schiffer's book (Schiffer 1987).

Lewis is well known for his "convention of truthfulness and trust in L," his answer to the question of what makes a particular language the language of a given community, a question that connects a language as an abstract object to language as used by its speakers (Lewis 1969, 1975a). That solution was not uncontroversial and he discussed various critiques of it in "Meaning without Use" (Lewis 1992). He has answers to most of the main critiques that have been offered, but found a new critique of his own, so in that paper he modifies some of his ideas from the two earlier works in ways that solve problems raised for earlier versions of his convention of truthfulness and trust. But that later

paper is not part of the "main canon" of Lewis's works that formal semanticists all know, and I will not discuss it here.

21.4 Possible Worlds, Counterfactuals, Modality, Counterparts

Most working linguists are not directly concerned with modal realism, and are quite willing to let the philosophers fight it out. And I will likewise leave commentary on Lewis's modal realism to others. But I should note that linguists have found the notion of counterparts useful in semantics. And not only in formal semantics: George Lakoff was the first linguist to pay attention to Lewis's counterpart theory (Lewis 1968, 1971), and he appealed to it (Lakoff 1968) in developing interesting new ideas about the distribution of reflexive and non-reflexive pronouns in some kinds of "split-identity" sentences, most famously in an example due to James McCawley:

(8) I dreamed that I was Brigitte Bardot and that I kissed me.

Lakoff was not entering into philosophical debates about whether an individual can exist in more than one world or not. His ideas may not have been directly related to Lewis's actual ideas so much as to problems of "deferred reference," as when I refer to a picture of myself in a photo saying "That's me," or in wax museum situations, or McCawley's dream scenario, etc. Similar issues were taken up and extended from picture-contexts to belief-contexts by Jackendoff (1975), also in ways not very closely related to Lewis's own ideas.

More serious influence of Lewis's modal realism and his ideas about possible worlds and modality can be seen work on propositional attitude contexts, and in ongoing debates, e.g. between the Lewis camp and the Kripke camp; I will not venture into those debates, which are discussed in other contributions. The work of the linguist Angelika Kratzer on conditionals (Kratzer 1979, 1986, 1989) has drawn heavily on Lewis's work on conditionals and conditional probability and also on his counterpart theory. Linguists learned a great deal about modality, conditionals, and the usefulness of possible worlds semantics from Lewis's and Stalnaker's work on counterfactual conditionals. I will not discuss conditionals in this paper; see Stalnaker's contribution to this volume (chapter 26).

21.5 Important Particular Ideas

In this section I mention a number of specific ideas of Lewis's that have been important for linguists, particularly but not only in the development of formal semantics.

21.5.1 Categorial Grammar and Function-Argument Structure

Linguists like me who learned about categorial grammar in the context of the Chomsky hierarchy learned only about its syntactic side, for syntacticians, categorial grammar syntax lost interest as soon as it was proven weakly equivalent to context-free phrase structure grammar. Lewis showed in GS its usefulness for compositional semantics, and the beauty of categorial grammar's systematic correspondence between syntactic categories and semantic types. Lewis's GS includes a beautiful pedagogical introduction to the key Fregean idea that function-argument application is a fundamental means of semantically composing parts into larger wholes. This idea was appreciated and developed by Curry (1930, 1961), but linguists mainly learned about the semantic

significance of categorial grammar from Lewis, who proposed a transformational syntax with a categorial deep structure and meaning-preserving transformations, and from Montague (1973, hereafter PTQ), who used categorial grammar in designing his syntactic categories and their correspondence to semantic types, although he did not restrict his syntax to the rules countenanced by categorial grammar.

Lewis proposed a categorial grammar for deep structure, while leaving open many choices. He argued that even requiring a categorial grammar is not a heavy constraint, and that it is not hard to find one that will resemble your favorite base component. In GS he provides a rich fragment, with many hints about extensions and alternatives; he shows his abilities as a linguist in his treatment of such phenomena as base-generated passive sentences and of reflexive pronouns.

One interesting possibility that he suggests: "Our system of meanings may serve, in effect, as a universal base for categorially based transformational grammars" (1970, 35). We may take the meanings as given, and regard a grammar as a certain way to encode meanings: the *representing relation* determined by the grammar. "If we were concerned with nothing but transformation-free categorial grammars, we could take a grammar to consist of nothing but a lexicon" (35): this is an early statement of an idea that has subsequently emerged in a number of contexts.

Linguists did indeed become more interested in exploring the possibility of making serious use of a categorial grammar framework once the semantic side of categorial grammars became clearer, and various versions of categorial grammar are being actively pursued (see, for instance, Bach, Oerhle and Wheeler 1987; Jacobson 1996; Dowty 2003; Steedman and Baldridge 2011).

In this context Lewis raised another methodological point that has played an important role in the development of formal grammar, one with respect to which I think his point of view has not prevailed. On page 29 of GS we find one of his famous quotes: "I promised simplicity; I deliver functions from functions from functions to functions to functions from functions to functions." And immediately after that remark he notes that in some cases, such as for extensional adjectives, whose extensions could simply be sets, it would be possible to find simpler intensions, "but at an exorbitant cost: we would have to give up the uniform function-and-arguments form for semantic projection rules." Actually, in this regard, as in the case of Montague's uniform category-to-type correspondence, linguists have found empirical arguments for sometimes after all favoring an intuitively simpler analysis over a more general one. Against Montague's uniform treatment, Partee and Rooth (1983) give arguments from generalized conjunction for treating extensional transitive verbs as being of the simpler type <e,<e,t>> while leaving intensional ones like *seek* in the category that Montague had assigned to all transitive verbs. And, against Lewis's contention that it would be a mistake to give up the uniformity of function-argument structure to treat extensional adjectives as just denoting sets, arguments from the logic of modifiers by Parsons (1990) and by Landman (2000) showed that sets (of entities for extensional adjectives, of events for extensional adverbials) actually give the right results in accounting for entailments involving permutation of modifiers and/or omission of a modifier. So while function-argument structure still plays a central role in formal semantics, few would now argue the uniform treatment that Lewis argued for.

One more note about functions: Lewis proposes a possible account of selectional restrictions via conditions on definedness of functions. If the argument of a function is not of the right sort, the function simply cannot apply. This idea indeed became widespread; see the work of Heim on presupposition within her file-change semantics (Heim 1983), and the textbook (Heim and Kratzer 1998).

21.5.2 Generalized Quantifiers

Lewis in GS treats noun phrases like *every pig, some pig* as generalized quantifiers, of semantic type <<e,t>,t>. He made that innovation at approximately the same time as Montague in PTQ. It's not

absolutely clear who had the idea first, and whether one of them got it from the other or whether they thought of it independently. I've written about this (Partee 2013), where I report on evidence that seems to point to Montague having had the idea first.

Lewis differs from Montague in not applying the generalized quantifier to proper names. He shows how we might treat names as category N, and non-simple NPs as category S/(S/N). He also shows how to give two base structures for "Every boy loves some girl" by using binding. He also suggests the possibility of in situ generalized quantifier direct objects for "Lothario seeks a girl", offering as an alternative a transformational derivation from "Lothario strives to find a girl."

He takes some time to explain binding, binders (basically lambda abstractors, but without that terminology, and illustrated only for the first-order case), and variables. And he also takes pains to argue for the separation of quantification and variable binding. He shows, for instance, that if one treats *something* as a generalized quantifier, no binding is needed in sentences with no anaphora and no scope ambiguity, like *Something grunts*. Many of the same points can be made about the treatment in Montague's PTQ, but Lewis explains it in terms accessible to anyone who knows a little first-order logic, whereas Montague offers very little help to the reader.

Lewis recognizes that those who prize generalization might want to avoid having two different categories for NPs, N for *Porky* and S/(S/N) for *every pig*. And he shows two ways one might go to achieve a uniform category and type assignment for NPs.

The first way is identical to what Montague did in PTQ, treating all NPs as generalized quantifiers; since there is no acknowledgement by either to the other, I would assume that these were independent discoveries. On this approach he puts Porky* into category S/(S/N), and has it take the verb phrase as argument.

The other way, which is less familiar, "assimilates quantificational phrases to names" by interpreting names as denoting *characters*, where a *character* is a set of properties. (He cites Mates (1968) and Montague (1970a) as well as a 1969 talk by Montague.) There are *individual characters* (a maximal compatible set of properties) and others (he says *generic* but that's not a good term.) For this alternative he revises the set of basic categories to S, NP, and C (common noun); extensions for common nouns are sets of individual characters. Quantifiers become NP/C. This approach never caught on anywhere in the linguistic literature, so far as I know. The theory of generalized quantifiers, on the other hand, is now a thriving subfield of formal semantics.

21.5.3 Adverbs of Quantification

Lewis (1975b) treated "adverbs of quantification," like the adverbs *usually* and *always*, in sentences like (9) and (10).

(9) A quadratic equation usually has two different solutions.
(10) A man who owns a donkey always beats it now and then.

The response to Lewis's idea that such adverbs were "unselective quantifiers," and that *if*-clauses might often be best understood as restrictors on such quantifiers, was not immediate. But after the work of Heim (1982) on "file change semantics" and of Kamp (1981) on discourse representation theory, that idea turned out to be extremely fruitful, and it has led to the idea that indefinites like *a donkey* never (or perhaps rarely) have any quantificational force of their own, and also to lasting revisions in the semantics of *if*-clauses, of donkey-sentences, of unselective binding, and of discourse anaphora. This family of ideas has been the starting point of decades of work in semantics (and syntax), still going on.

21.5.4 Contexts and Indices

A great deal of the evolution of thinking about what meanings should do and finding things that can do that has involved "contexts" and "indices of evaluation." This is too big a topic to do justice to here. Lewis was an important figure in that evolution. In his early GS paper, he proposes to have Carnapian intensions as functions from indices to extensions, and follows the suggestion (Montague 1968; Scott 1970) of letting indices be "n-tuples of the various items besides meanings that may enter into determining extensions" (1970, 24). As a first suggestion, he proposes an octuple containing a possible world, a time, a place, a speaker, an addressee, a (possibly empty) set of things capable of being pointed at (the *indicated-objects* coordinate), a (possibly empty) stretch of discourse (the *previous discourse* coordinate), and an assignment of values to variables. In an appendix to GS, he expresses concerns that it may be hard to identify all the aspects of context that might enter into determining extensions. Cresswell (1972) objected, and Lewis subsequently agreed, that when it comes to contexts, it is impossible to specify any discrete number of aspects of context that will always be sufficient: contexts should be taken holistically, and then one can define "speaker of context c" etc. for indefinitely many properties of contexts.

In later works, including Lewis (1974, 1980), he was part of ongoing explorations concerning the distinction between those parameters, finite in number, that are compositionally manipulated ("index"), as the time and world indices are manipulated in the interpretation of sentences with tense and modal embeddings, and those that are parts of the context and seem not to be manipulated compositionally – like the speaker and hearer parameters. Kaplan had one way of regimenting that distinction with his notion of *character* (Kaplan 1979); Stalnaker (1970) had another. Lewis (1974, 1980) took pains to show what kinds of different options could be explored, how to translate one approach into another, and where different choices did and didn't make a difference.

In a postscript to GS (Lewis 1983) he summarizes the importance of distinguishing between context and index, and suggests taking a context to be a package of a world, a time, and a speaker, "Or, as I would prefer, with all three in one: a world-bound time-slice of a (potential) speaker" (1983, 230). He notes in a footnote that these "contexts" are the same as the "subjects" that self-ascribe properties in his paper "Attitudes *De Dicto* and *De Se*" (Lewis 1979a).

That paper, "Attitudes *De Dicto* and *De Se*," had a big impact in linguistics, because it soon became clear that the kind of "identity" involved in infinitival control was stronger than coreference, stronger than bound variable identity, and really must be analyzed as *de se* anaphora. Early transformational grammar had derived a sentence like (11) from an underlying structure something like (12); but quantificational examples like (13) showed that syntactic identity was certainly not the right relation to posit between the subjects of the two clauses.

(11) John tried to win.
(12) John tried [John to win]
(13) Every man tried to win.

So later proposals posited something like variable binding: *x tried* [*x win*], but even that was not strong enough to capture the *de se*-ness of the relation. It was Chierchia (1989b) who showed how Lewis's work on *de se* attitudes could be fruitfully applied to the analysis of controlled infinitives. The infinitival phrase should be property-denoting, rather than analyzed as a clause; and the control verb (*try* in this example) should have the *de se* ascription of that property to the subject built in as part of its lexical semantics. A range of interesting linguistic consequences follow. (See also chapter 25 by Richard Holton on *de se* attitudes in this volume.)

21.6 Bridging Philosophy and Linguistics

After Richard Montague's death in 1971, several linguists and philosophers began to work on extending "Montague grammar" and connecting it to work in linguistics. David Lewis, while not directly doing "Montague grammar," was doing work very close in spirit and practice to what Montague had been doing, and with more attention to what linguists had to contribute.

He was a part of the first international conference on formal semantics of natural language, organized by Edward Keenan at the University of Cambridge in April 1973, a conference that included approximately equal numbers of linguists and philosophers, at which he presented his "Adverbs of Quantification," which was published in the conference volume in 1975. As I mentioned in section 21.5, that work became highly influential among semanticists.

David had moved to Princeton in 1970. I moved to the University of Massachusetts, Amherst in 1972, as did the philosopher Terry Parsons, another early contributor to the development of formal semantics and my close colleague here in the 1970s, where he, Emmon Bach, and I did joint research and teaching with both linguistics and philosophy graduate students. Together we trained a number of the young formal semanticists of the next generation. David had encouraged me to consider moving to Princeton instead, and I did consider it seriously, largely because of the appeal of having David as a colleague. But it was clear that Princeton did not want to have more than two or three linguists, and was not going to be building up a PhD program in linguistics, so the attraction of being connected to David and that excellent Philosophy Department could not outweigh the prospect of helping to build up a new linguistics PhD program that could have formal semantics as one of its specializations, with another fine philosopher of language as a close colleague.

In the summer of 1974, UMass Amherst hosted the then annual Linguistic Institute of the Linguistic Society of America, an eight-week international event with courses, workshops, and special lectures. I was put in charge of the semantics and philosophy of language part of the Institute, one of its special foci. We had a large number of courses in those areas, and two workshops funded by the Mathematical Social Sciences Board, one that I organized on "The Syntax and Semantics of Non-Extensional Constructions," and one that Emmon Bach organized on non-transformational grammars, including categorial grammars. The faculty who taught courses in semantics and philosophy of language were a subset of those who participated in my workshop – the workshop made it possible to bring in a large number of scholars to teach one course and do research in the workshop. The invited workshop participants included Rich Thomason, Bob Stalnaker, David Lewis, Emmon Bach, Terry Parsons, David Dowty, Ray Jackendoff, Janet Fodor, Ed Keenan, Hans Kamp, Lauri Karttunen, Michael Bennett, Enrique Delacruz, and two graduate students, Anil Gupta (Thomason's student) and Robin Cooper (mine). David Lewis was a very valuable contributor to that group, because he was not only an expert on intensionality (as well as hyperintensionality) but also understood the linguists and our issues very well. I think his paper "Truth in Fiction" was drafted around that time, and we probably discussed it in the workshop, since the problem of the semantics of sentences in fiction was one of our topics; Terry Parsons was also working on it at the time. And I think his paper "Whether Report" may also have been circulating in draft form by then.

Around that time I started putting together the first volume of papers on "Montague grammar" (Partee 1976), which includes contributions by many of the participants in that 1974 workshop (Lewis, Partee, Thomason, Bennett, Delacruz, Dowty, Cooper, and Parsons) and some others (Hamblin, Cresswell, Rodman, and Siegel). I invited David to contribute an original article to the volume, but he declined, saying that philosophy of language was only one of his interests and he was just then more focused on working on some problems in metaphysics. But he kindly gave me permission to

reprint his "General Semantics," which appeared as the first paper in the volume and served as an ideal introduction to the aims and methods of formal semantics.

I was at several more conferences in the 1970s that involved linguists and philosophers at which David played an important and constructive role. One was at SUNY Albany in April 1977, organized by the philosopher Steven Davis and the linguist Marianne Mithun, called Linguistics, Philosophy, and Montague Grammar (Davis and Mithun 1979). David doesn't have a paper in the volume, but I am quite sure he was there, along with philosophers Rich Thomason, Terry Parsons, Michael Bennett, James Waldo, and the Amsterdam team of Jeroen Groenendijk and Martin Stokhof (at their first conference in the USA), and linguists Emmon Bach, me, Jim McCawley, David Dowty, and Muffy Siegel. At that time I was beginning to voice my worries about how Montague-style semantics could be "in the head" as a Chomskyan view of linguistics would demand, and David's comments about separating what the semantics is from how our knowledge of it is to be described made a big and helpful impression on me.

Another memorable conference was in Uppsala in June 1977, on the occasion of the 500th anniversary of Uppsala University, on Philosophy and Grammar. That conference did bring linguists and philosophers together, but it could not be categorized as a formal semantics conference. I was taken quite by surprise in the discussion after my paper, which was one of the first. I had been nervous to have Quine and David Lewis sitting in the front rows, but at least I had some idea of their positions on the semantic issues I was discussing. The surprise came when some of the many Scandinavian logicians and philosophers there such as Per Martin-Löf and Dag Prawitz responded to my paper by challenging some of the basic Fregean and Tarskian foundational principles on which formal semantics was built. I had known nothing about them and their work and was ill prepared for their questions. David was very helpful in the discussion period; he was able to bridge our mutual noncomprehension with clear and constructive comments. Others at the conference included Quine, Hintikka, Dagfinn Føllesdal, Arne Naess, James McCawley, Stig Kanger, and Sven Öhman. David's own paper at that conference was "Index, Context, and Content" (Lewis 1980). An earlier version had been presented to the Vacation Summer School in Logic at the Victoria University of Wellington in August 1976. That paper, which I described briefly in section 21.5.4, was well received and generated interesting discussion. The conference, despite the theoretical differences among participants, was a very friendly one. Quine's wife had come with him, and some other spouses as well, and we were treated to a feast in a castle overlooking the university. In such a setting intellectual disagreement was just stimulating.

David's "Scorekeeping in a Language Game," central to chapter 22 in this volume by Craige Roberts, was presented in several places; I heard it at the Konstanz conference of September 1978 on Semantics from Different Points of View (Bäuerle, Egli and von Stechow 1979). That was, I believe, the first international formal (and other) semantics conference in Germany, organized by Arnim von Stechow and two colleagues. Other participants included David's (and Arnim's) good friend Max Cresswell, Rainer Bäuerle, Hans Kamp, me, Angelika Kratzer, and Irene Heim. The conference included some other linguists, philosophers, and psycholinguists, whose contributions were not in formal semantics, but the core was formal semantics, and it was an occasion when significant new and lasting contacts were made.

That paper of David's was also published in the *Journal of Philosophical Logic* in 1979. An early version of this paper had been presented to the Vacation School in Logic at Victoria University of Wellington in August 1976, and a near-final version at a workshop on pragmatics and conditionals at the University of Western Ontario in May 1978.

So indeed David Lewis was a central figure in that crucial decade of the 1970s, when linguists and philosophers interacted productively to develop formal semantics, to integrate it with linguistic work, to broaden its scope to address a range of challenging problems in natural language semantics

and the emerging formal pragmatics. In fact most of his work on philosophy of language and semantics dates from the 1970s; it is fortunate for the field that he was most active in that area just then. I've mentioned four conferences that we both participated in (in addition to the Linguistic Institute in Amherst in the summer of 1974), and all four of the corresponding conference papers have become classics. "General Semantics," from the La Jolla conference of 1969, was anthologized in the first collection of papers on Montague grammar, and all semanticists know it. "Adverbs of Quantification," from Keenan's 1973 conference in Cambridge, is included in a volume of essential readings in formal semantics (Portner and Partee 2002), as is "Scorekeeping in a Language Game," from the Konstanz conference of 1978. I don't know whether "Index, Context, and Content," from the 1977 Uppsala conference, has been anthologized in any semantics collection, but it is almost as well known, as are several of his other papers from the 1970s that I have discussed in this chapter. (Many of his papers are anthologized in philosophy of language collections, of course.) Philosophy of language may not have been his first love, and his work in that area may not be what philosophers consider his greatest work, but for formal semanticists he clearly ranks as one of the greatest contributors to the creation of our field.

Acknowledgments

Thanks to Thomas Blanchard for providing me with electronic copies of all of David Lewis's language-related papers while I was in Moscow, and for assistance with formatting. I thank David Kaplan and Stephanie Lewis for granting me interviews in connection with my current book project on the history of formal semantics; I have drawn on those interviews in section 21.1. And thanks to Stephanie Lewis for reading the penultimate draft of this paper. I wish I could thank David Lewis. I have tried to avoid hagiography, but I cannot pretend to conceal my admiration for David Lewis and for his work.

Notes

1 I'm not the only one to have a first impression of David as shy and socially uncomfortable. David Kaplan reports (unpublished interview with the author) that when the hiring committee from UCLA went to the December APA meetings to interview candidates, they looked around for David Lewis and found him standing behind a palm tree in a corner. Kaplan says he was nervous and shy, awkward at socializing; but very smart; hiring him felt like a shot in the dark, but they never regretted it. And he never seemed shy in situations he was comfortable in, like discussing philosophy.

2 It was there that I first heard of Chomsky, and learned that he was about to start a PhD program in linguistics at MIT. My own path into linguistics, with philosophy as a continuing second interest, is described in Partee 2004 and 2005.

3 Unpublished interview with the author.

4 David Kaplan reports (unpublished interview with the author) that David and Steffi never felt totally comfortable in Los Angeles; they would say things like "The trees don't look right." Steffi (pers. comm.) confirms that.

5 Addressee unknown. Letter courtesy of Stephanie Lewis.

6 From a letter from Bar-Hillel to Montague, November 20, 1967 (in the Montague archives, UCLA), written after receipt of one of Montague's pragmatics papers: "It will doubtless be a considerable contribution to the field, though I remain perfectly convinced that without taking into account the recent achievements in theoretical linguistics, your contribution will remain one-sided." Similar sentiments are expressed in some of Staal's letters to Montague; both Staal and Bar-Hillel made efforts to bring Montague and Chomsky into some kind of constructive contact, without success.

7 Just a reminder: many of the papers discussed here, including this one and GS, circulated as manuscripts for years before publication, so who influenced whom cannot be inferred from publication dates.

References

Bach, E., Oehrle, R. & Wheeler, D. (Eds.) (1987). *Categorial grammars and natural language structures*. Dordrecht: Reidel.

Bäuerle, R., Egli, U. & von Stechow, A. (Eds.) (1979). *Semantics from different points of view*. Berlin: Springer.

Carnap, R. (1947). *Meaning and necessity*. Chicago: University of Chicago Press.

Chierchia, G. (1989a). Structured meanings, thematic roles and control. In G. Chierchia, B. Partee & R. Turner (Eds.), *Properties, types, and meaning*, pp. 131–166. Dordrecht: Kluwer.

Chierchia, G. (1989b). Anaphora and attitudes *de se*. In R. Bartsch, J. van Benthem & P. van Emde Boas (Eds.), *Semantics and contextual expression*. Dordrecht: Foris.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Cresswell, M.J. (1972). The world is everything that is the case. *Australasian Journal of Philosophy*, 50, 1–13.

Cresswell, M.J. (1973). *Logics and languages*. London: Methuen.

Cresswell, M.J. (1985). *Structured meanings: the semantics of propositional attitudes*. Cambridge, MA: MIT Press.

Curry, H.B. (1930). Grundlagen der kombinatorischen logik. *American Journal of Math*, 52, 509–36, 789–934.

Curry, H.B. (1961). Some logical aspects of grammatical structure. In R. Jakobson (Ed.), *Structure of language and its mathematical aspects*, pp. 56–68. Providence, RI: American Mathematical Society.

Davis, S. & Mithun, M. (Eds.) (1979). *Linguistics, philosophy, and Montague grammar*. Austin and London: University of Texas Press.

Dowty, D. (1979). *Word meaning and Montague grammar. The semantics of verbs and times in Generative Semantics and in Montague's PTQ*. Synthese Language Library. Dordrecht: Reidel.

Dowty, D. (2003). The dual analysis of adjuncts/complements in categorial grammar. In E. Lang, C. Maienborn & C. Fabricius-Hansen (Eds.), *Modifying adjuncts*, 33–66. Berlin and New York: Mouton de Gruyter.

Harris, R.A. (1993). *The linguistics wars*. New York and Oxford: Oxford University Press.

Heim, I. (1982). The semantics of definite and indefinite noun phrases. PhD dissertation, University of Massachusetts; published 1989, New York: Garland.

Heim, I. (1983). On the projection problem for presuppositions. In M. Barlow, D. Flickinger & M. Wescoat (Eds.), *WCCFL 2: second annual West Coast Conference on Formal Linguistics*, pp. 114–25 Stanford, CA: Stanford University Press.

Heim, I. & Kratzer, A. (1998). *Semantics in generative grammar*. Oxford: Blackwell.

Jackendoff, R. (1975). On belief-contexts. *Linguistic Inquiry*, 6, 53–93.

Jackendoff, R. (1996). Semantics and cognition. In S. Lappin (Ed.), *The handbook of contemporary semantic theory*, pp. 539–59. Oxford: Blackwell.

Jacobs, R.A. & Rosenbaum, P.S. (1968). *English transformational grammar*. Lexington, MA: Xerox College Publishing.

Jacobson, P. (1996). The syntax/semantics interface in categorial grammar. In S. Lappin (Ed.), *The handbook of contemporary semantic theory*, pp. 89–116. Oxford: Blackwell.

Kamp, H. (1981). A theory of truth and semantic representation. In J.A.G. Groenendijk, T.M.V. Janssen & M.B.J. Stokhof (Eds.), *Formal methods in the study of language*, Mathematical Centre Tracts 135, pp. 277–322. Amsterdam: Mathematical Centre.

Kaplan, D. (1979). On the logic of demonstratives. *Journal of Philosophical Logic*, 8, 81–98.

Katz, J.J. & Fodor, J.A. (1963). The structure of a semantic theory. *Language*, 39, 170–210.

Katz, J.J. & Postal, P.M. (1964). *An integrated theory of linguistic descriptions*. Cambridge, MA: MIT Press.

Kratzer, A. (1979). Conditional necessity and possibility. In R. Bäuerle, U. Egli & A. von Stechow (Eds.), *Semantics from different points of view*, pp. 117–47. Berlin: Springer.

Kratzer, A. (1986). Conditionals. In A.M. Farley, P. Farley & K.-E. McCullough (Eds.), *CLS 22: papers from the parasession on pragmatics & grammatical theory, the twenty second meeting of the Chicago Linguistic Society*, pp. 115–35. Chicago: Chicago Linguistic Society.

Kratzer, A. (1989). An investigation of the lumps of thought. *Linguistics and Philosophy*, 12, 607–53.

Lakoff, G. (1968). *Counterparts, or the problem of reference in transformational grammar*. Bloomington: Indiana University Linguistics Club.

Lakoff, G. (1971). On generative semantics. In D. Steinberg & L. Jakobovits (Eds.), *Semantics: an interdisciplinary reader in philosophy, linguistics, and psychology*, pp. 232–96. Cambridge: Cambridge University Press.

Lakoff, G. (1987). *Women, fire, and dangerous things*. Chicago: University of Chicago Press.

Landman, F. (2000). *Events and plurality: the Jerusalem Lectures: Studies in linguistics and philosophy*, Vol. 76. Dordrecht: Kluwer.

Lewis, C.I. (1944). The modes of meaning. *Philosophy and Phenomenological Research*, 4, 236–49.

Lewis, D. (1968). Counterpart theory and quantified modal logic. *Journal of Philosophy*, 65, 113–26.

Lewis, D. (1969). *Convention: a philosophical study*. Cambridge, MA: Harvard University Press.

Lewis, D. (1970). General semantics. *Synthese*, 22, 18–67.

Lewis, D. (1971). Counterparts of persons and their bodies. *Journal of Philosophy*, 68, 203–11.

Lewis, D. (1974). 'Tensions. In M. Munitz & P. Unger (Eds.), *Semantics and philosophy*. New York: New York University Press.

Lewis, D. (1975a). Languages and language. In K. Gunderson (Ed.), *Language, mind, and knowledge*, pp. 3–35. Minneapolis: University of Minnesota Press.

Lewis, D. (1975b). Adverbs of quantification. In E.L. Keenan (Ed.), *Formal semantics of natural language*, pp. 3–15. Cambridge: Cambridge University Press.

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *Philosophical Review*, 85, 297–315.

Lewis, D. (1978). Truth in fiction. *American Philosophical Quarterly*, 15, 37–46.

Lewis, D. (1979a). Attitudes *de dicto* and *de se*. *Philosophical Review*, 88, 513–43.

Lewis, D. (1979b). Scorekeeping in a language game. In R. Bäuerle, U. Egli & A. von Stechow (Eds.), *Semantics from different points of view*. Berlin: Springer.

Lewis, D. (1980). Index, context, and content. In S. Kanger & S. Öhman (Eds.), *Philosophy and grammar*, pp. 79–100. Dordrecht: Reidel.

Lewis, D. (1983). Postscripts to "General semantics." In D. Lewis, *Philosophical papers*, Vol. I, pp. 230–2. Oxford: Oxford University Press.

Lewis, D. (1992). Meaning without use: reply to Hawthorne. *Australasian Journal of Philosophy*, 70(1), 106–10.

Mates, B. (1968). Leibniz on possible worlds. In B. van Rootselaar & J.F. Staal (Eds.), *Logic, methodology, and philosophy of science III*. Amsterdam: North–Holland.

McCawley, J.D. (1968). Concerning the base component of a transformational grammar. *Foundations of Language*, 4, 243–69.

Montague, R. (1968). Pragmatics. In R. Klibanski (Ed.), *Contemporary philosophy*, pp. 102–21. Florence: La Nuova Italia.

Montague, R. (1970a). Universal grammar. *Theoria*, 36, 373–98.

Montague, R. (1970b). English as a formal language. In B. Visentini et al. (Eds.), *Linguaggi nella società e nella tecnica*, pp. 189–224. Milan: Edizioni di Comunità.

Montague, R. (1973). The proper treatment of quantification in ordinary English. In K.J.J. Hintikka, J.M.E. Moravcsik & P. Suppes (Eds.), *Approaches to natural language*, pp. 221–42. Dordrecht: Reidel.

Newmeyer, F.J. (1980). *Linguistic theory in America: the first quarter century of transformational generative grammar*. New York: Academic Press.

Parsons, T. (1990). *Events in the semantics of English. a study in subatomic semantics*. Cambridge, MA: MIT Press.

Partee, B.H. (1970). Negation, conjunction, and quantification: syntax vs. semantics. *Foundations of Language*, 6, 153–65.

Partee, B.H. (Ed.) (1976). *Montague grammar*. New York: Academic Press.

Partee, B.H. (1977). Possible worlds semantics and linguistic theory. *The Monist* 60, 303–26.

Partee, B.H. (1979). Semantics – mathematics or psychology? In R. Bäuerle, U. Egli & A. von Stechow (Eds.), *Semantics from different points of view*, pp. 1–14. Berlin: Springer.

Partee, B.H. (1988). Semantic facts and psychological facts. *Mind & Language*, 3, 43–52.

Partee, B.H. (2004). Reflections of a formal semanticist. In B. Partee, *Compositionality in formal semantics: selected papers by Barbara H. Partee*, pp. 1–25. Oxford: Blackwell.

Partee, B.H. (2005). Reflections of a formal semanticist as of Feb 2005, http://people.umass.edu/partee/docs/BHP_Essay_Feb05.pdf, accessed June 3, 2014.

Partee, B.H. (2013). Montague's "linguistic" work: Motivations, trajectory, attitudes. In E. Chemla, V. Homer & G. Winterstein (Eds.), *Proceedings of Sinn und Bedeutung 17, September 8–10 2012*, pp. 427–53. Paris: ENS.

Partee, B.H. & Rooth, M. (1983). Generalized conjunction and type ambiguity. In R. Bäuerle, C. Schwarze & A. von Stechow (Eds.), *Meaning, use, and interpretation of language*, pp. 361–83. Berlin: De Gruyter.

Peters Jr., P.S. & Ritchie, R.W. (1973). On the generative power of transformational grammars. *Information Sciences*, 6, 49–83.

Pietroski, P. (2003). The character of natural language semantics. In A. Barber (Ed.), *Epistemology of Language*, pp. 217–56. Oxford: Oxford University Press.

Portner, P. & Partee, B.H. (Eds.) (2002). *Formal semantics: the essential readings*. Oxford: Blackwell.

Putnam, H. (1978). *Meaning and the moral sciences*. Boston, London, and Henley: Routledge & Kegan Paul.

Ross, J.R. (1970). On declarative sentences. In R.A. Jacobs & P.S. Rosenbaum (Eds.), *Readings in English transformational grammar*, pp. 222–72. Waltham, MA: Ginn & Co.

Schiffer, S. (1987). *Remnants of meaning*. Cambridge, MA: MIT Press.

Scott, D. (1970). Advice on modal logic. In K. Lambert (Ed.), *Philosophical problems in logic: some recent developments*, pp. 143–74. Dordrecht: Reidel.

Stalnaker, R. (1970). Pragmatics. *Synthese*, 22, 272–89.

Steedman, M. & Baldridge, J. (2011). Combinatory categorial grammar. In R.D. Borsley & K. Börjars (Eds.), *Non-transformational syntax: formal and explicit models of grammar*, pp. 181–224. Oxford: Wiley-Blackwell.

Stenius, E. (1967). Mood and language-game. *Synthese*, 17, 254–74.

van Lambalgen, M. & Hamm, F. (2005). *The proper treatment of events: explorations in semantics*. Oxford: Blackwell.

von Stechow, A. (1984). Structured propositions and essential indexicals. In F. Landman & F. Veltman (Eds.), *Varieties of formal semantics*, pp. 385–403. Dordrecht: Foris.

Weatherson, B. (2010). David Lewis. In E.N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2010 edition), http://plato.stanford.edu/archives/sum2010/entries/david-lewis/, accessed June 12, 2014.

22

Accommodation in a Language Game

CRAIGE ROBERTS

His argument true, his tone light.

Seamus Heaney, "Saint Francis and the Birds"

22.1 Introduction

The conventional content of a typical linguistic utterance underdetermines the speaker's intended meaning. In cases involving anaphora, indexicals, and ellipses, the problem for a compositional theory of semantic interpretation is generally more or less resolved with the use of contextual indices, their use triggered by elements of the conventional content of the utterance. Where there are no overt triggers, theorists posit covert variables in the logical form of the utterances in question (see the papers and references in section 7 of Partee 2004). But it's one thing to deal in this way with cases where a free variable will suffice, treating the context-sensitivity as effectively indexical. It's quite another to address phenomena like conversational implicature which intrinsically involve abductive inference based on rich contextual information. Most often, such contributions to intended meaning are shuttled off to a post-semantic enrichment phase: Gricean icing – after the fact – on a truth conditional cake. But when we look carefully at a wider variety of context-sensitive factors and how they interact with truth conditional interpretation, this neat division becomes problematic. So is it possible to say anything interesting, in the scientific sense of *falsifiable*, about such contextual factors in interpretation?

David Lewis's work in this area has been very influential not only in the philosophy of language, but also in the linguistic literature on formal semantics and pragmatics. "Scorekeeping in a Language Game" (1979) is the article by Lewis which is most often cited in the linguistic literature. In it, he tackles presupposition, vagueness, performatives, and other messy phenomena that don't fit readily into the neat indexical box. And against the backdrop of his earlier work on language games, linguistic convention and compositional syntax/semantics, he presents two new theses which are intended to contribute to a framework for addressing such sticky issues in the study of semantics:

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

- Linguistic context is best modeled as a *scoreboard* in the language game.
- Unlike in competitive games, *accommodation* is a regular component of the dynamics of score change in the language game: "conversational score . . . tend[s] to evolve in such a way as is required to make whatever occurs count as correct play." (1979, 346)

Lewis says little about the actual content of the scoreboard, beyond a few hints at possible elements of the score: "The rules specifying the kinematics of score thereby specify the role of a scoreboard; the scoreboard is whatever best fills this role; and the score is whatever this scoreboard registers" (1979, 346). He says nothing about how the score might facilitate or constrain accommodation. But I think we can glean a great deal from the scoreboard metaphor itself, given how his earlier work used game theory to characterize linguistic convention.

In that work, the language game is essentially a cooperative endeavor (Lewis 1969), whose participants have common goals – roughly, the accurate sharing of information. Hence, they are motivated to behave in an accommodating fashion:

> If at time *t* something is said that requires component *sn* of conversational score to have a value in the range *r* if what is said is to be true, or otherwise acceptable; and if *sn* does not have a value in the range *r* just before *t*; and if such-and-such further conditions hold; then at *t* the score-component *sn* takes some value in the range *r*. (Lewis 1979, 347)

This first, general definition is subsequently modified to yield a number of more specific types of accommodation, including presupposition accommodation (below).

One of the most straightforward conversational circumstances in which accommodation commonly occurs is that involving a so-called *performative speech act*, as in the following (Lewis 1979, 355, with my glosses in square brackets):

(1) I hereby name this ship the *Generalissimo Stalin*.

[compositionally asserted: 'the speaker names this ship the *Generalissimo Stalin*'
verified by proper performance: 'this ship is named the *Generalissimo Stalin*']

(2) I now pronounce you man and wife.

[compositionally asserted: 'the speaker pronounces the addressees to be man and wife'
verified by proper performance: 'the addressees are man and wife']

Utterance of such a performative is a certain type of socially licensed act. When various preconditions hold – for example, the locution is performed by a speaker who is empowered by the appropriate authorities to perform the speech act in question, there is a marriage license, etc. – then as a consequence of the performance another proposition is verified, in the etymological sense of 'made true'. The verified proposition is neither the compositional meaning of the utterance (what it asserts), nor the usual secondary contribution of an assertion to the interlocutors' common ground (the proposition that the act itself was performed; Stalnaker 1979), but the proposition that the addressees are married. Note that the felicity of the act itself seems to require that this proposition was *not* true prior to the act; so this is not accommodation of a presupposition. But by virtue of the authority vested in the speaker, the act *makes this proposition true in the actual world*. Since that is so, the competent, cooperative interlocutor, when confident that the preconditions of authorized performance of the act are satisfied, then accommodates the truth of the conventional result to the conversational score.

Lewis presents a number of other types of circumstances in which accommodation is called for in conversation, including domain restriction (instantiated by his "relative modality" case), and the interpretation of definite descriptions. But surely the most influential idea in this paper is the application of accommodation to cases involving presupposition:

> *rule of accommodation for presupposition*:
> If at time t something is said that requires presupposition P to be acceptable, and if P is not presupposed just before t, then – *ceteris paribus* and within certain limits – presupposition P comes into existence at t. (1979, 340)

Presupposition accommodation bears directly on a central problem for the classical theory of presupposition and of presupposition projection due to linguist Lauri Karttunen (1973, 1974) and philosopher Robert Stalnaker (1973), the general view of presupposition which Lewis adopts (358, fn 1). On that account, a speaker in making a given utterance presupposes proposition P just in case the felicity of the utterance necessitates that P be entailed by the interlocutors' common ground at the time of utterance. For convenience, let's informally call this a requirement that P be contextually *given* (at the time of utterance). This is intended to explain the fact that when a speaker uses a factive verb like *regret* in (3), he seems to presume the truth of the proposition denoted by the complement ('Susan bought a ferret') rather than asserting it.

(3) Susan regrets that she bought a ferret.

Among other virtues, this approach offers a simple, intuitive explanation of the problem of presupposition projection, wherein the presumption of the truth of the complement persists ("projects") in a number of embedding contexts where other aspects of the conventional content of the clause do not survive: under negation, interrogation, a modal auxiliary, or in the antecedent of a conditional. Hence, in (4) the proposition that Susan bought a ferret seems to be implicated in each variant of (3), despite the fact that the proposition that Susan regrets something does not survive the embedding.

(4) Variants of (3) under embedding:

negation:	Susan doesn't regret that she bought a ferret.
interrogation:	Does Susan regret that she bought a ferret?
modal auxiliary:	Susan may regret that she bought a ferret.
conditional antecedent:	If Susan regrets that she bought a ferret, she can sell it.

If felicitous utterance of these forms requires that the complement proposition 'Susan bought a ferret' already be contextually given, this would explain why it seems to project from under the scope of these operators: If the utterance of one of the forms in (3)/(4) is felicitous, the complement just recapitulates something we already took to be true. So in (4), the negation, interrogation, or hypotheticality can only felicitously pertain to Susan's regret.

But there is a problem widely acknowledged in the literature. This theory appears to falter in accounting for examples wherein an utterance containing content that projects seems to be perfectly acceptable despite the fact that the projective content is clearly novel information in the context of utterance. With factive verbs like *regret* the complement is often new information, even explicitly so. Consider (5):

(5) I regret that I must inform you that Susan bought a ferret.

Here, use of *inform* implies that so far as the speaker knows the information conveyed is novel to the addressee, whereas *regret* supposedly presupposes that the addressee already knows that the speaker must so inform her. Thus, the theory predicts that a pragmatic contradiction should arise here. But none is attested by native speakers: The utterance is deemed felicitous in the context described, where the speaker's obligation is news, along with the information it pertains to.

I don't have the space here to discuss the virtues of the Karttunen/Stalnaker approach to presupposition (see Beaver 2001 for a detailed critical review). But its elegance has motivated many to attempt to save it in the face of such prima facie counterevidence by appeal to Lewis's accommodation. For example, the prominent theories of presupposition and presupposition projection due to Heim (1982, 1983) and van der Sandt (1999) make extensive use of this principle. But there is a growing literature debating the role and nature of accommodation (Thomason 1990; Thomason Stone and DeVault 2006; Abbott 2008; Gauker 2008; Stalnaker 2008; von Fintel 2008); and some of these authors (see also Simons *et al.* 2010) suggest that the high frequency and felicity of utterances which, according to these accounts of presupposition, require accommodation undercuts the foundations of this approach to explaining projection. It is my contention that all parties to this debate would benefit from taking a broader look at the phenomenon of accommodation and its role in discourse *in view of the nature of the language game*, and that this, in turn, sheds light on the nature of both linguistic presupposition and presupposition projection.

In this chapter, I'll focus on four questions which I take to be useful in understanding presupposition accommodation as Lewis defines it above: The first is a question about how we recognize that an utterance involves a presupposition. The first conjunct of the protasis of Lewis's conditional assumes that accommodation kicks in when something has been said that "requires presupposition P to be acceptable." But in the general case, it turns out to be non-trivial to say (a) how we recognize that something has been presupposed, and (b) how we retrieve *what* is presupposed. How *do* interlocutors retrieve P? We'll explore this in section 22.2.

The second question is about what it is to accommodate. We would expect an adequate answer to this question to shed light as well on the question of why interlocutors are inclined to do it. In section 22.3 we'll consider a characterization of accommodation due to Thomason (1990), and argue that it appropriately extends the range of phenomena taken to involve accommodation, while shedding light on the kinds of cases Lewis considered, and in particular on presupposition accommodation narrowly.

The third question has to do with the role of the scoreboard in accommodation. It seems reasonable to assume that in keeping with Lewis's general program for a natural language semantics, his scoreboard is intended to provide us with a model of the context of utterance with respect to which the proposition P is "acceptable" but "not presupposed" just before the utterance. In section 22.4, I'll discuss a theory of the scoreboard building on Lewis's characterization which crucially reflects the interlocutors' recognized goals and plans in the language game, and argue that this type of scoreboard plays a natural role in facilitating and constraining accommodation.

The fourth question has to do with Lewis's *ceteris paribus* condition. What are the limits on accommodation in discourse? When does infelicity result from presupposition failure, despite the cooperative intentions of the interlocutors? Again I will argue, in section 22.5, that some strong limits are a natural consequence of the nature of the language game and its scoreboard. Some conclusions are drawn in section 22.6.

The general strategy I adopt here is only partially exegetical. It is just as much an attempt at explication of the notions Lewis sketches in this paper, and an exploration of the general methodology for pragmatics it suggests. I'm not sure that Lewis himself would agree with all I have to say. But I believe the proposed strategy is consistent with the general program for semantics he lays out, and certainly with the direction in which contemporary research on semantics and pragmatics influenced by Lewis is proceeding.

22.2 Presupposition Recognition

> Presuppositions can be created or destroyed in the course of a conversation. This change is rule-governed, at least up to a point. (Lewis 1979, 339)

Even the most cooperative interlocutor can only accommodate something presupposed by a speaker if she recognizes (a) *that* something has been presupposed, and (b) *what* that presupposition is. The discussion of accommodation often seems to take it for granted that presupposition recognition has taken place, but several kinds of examples illustrate why this is not generally a reasonable assumption, and that the problem of presupposition recognition can be at least as challenging as that of presupposition accommodation.

A good deal of the literature on presupposition accommodation tends to focus on informative factive verbs and novel possessive descriptions, as in the following:

(6) We regret that children cannot accompany their parents to commencement exercises. (Karttunen 1974[1]; Gauker 2008)
(7) I can't come to the meeting – I have to pick up my cat at the veterinarian. (Stalnaker 1998; von Fintel 2008)

The underlined expressions in these examples have been taken to trigger presuppositions – the truth of the complement of *regret*, or with possessive *my* the existence (and possibly uniqueness) of an entity having the property denoted by the noun, in (7) a cat owned by the speaker. As the reader can verify, the factive implication in (6) does project in the classical contexts used to test projection, illustrated in (3) above. The same tests show that the existence implications triggered by possessive N(oun) P(hrase)s tend to project, as well. Since they trigger projective implications, *regret* and other factives, and possessive NPs are said to trigger presuppositions. Then, according to the classical theory due to Karttunen and Stalnaker, we would expect them to impose a condition on felicitous utterance: that of givenness in prior context, requiring prior knowledge on the part of the addressee(s).

Yet, as with (5) above, (6) can be quite felicitously, even politely used to knowingly *inform* the addressee(s) that children cannot accompany their parents, and (7) may be cooperatively uttered to a near-stranger who (as the speaker is aware) doesn't know that the speaker has any pets. Those who adopt classical presupposition theory then attempt to explain the very common informative use of such utterances by claiming that they involve accommodation. But if accommodation is so common, doesn't that undermine the theory of presupposition and projection whose proponents must appeal to it in so many cases? That is, if accommodation can save the day when presuppositions fail to be given in context, how can the classical theory be falsified?

The debate about accommodation focuses on examples like (6) and (7). Among the most prominent participants in the debate, Stalnaker (1998) and von Fintel (2008) have proposed ways of characterizing the timing of the requirement of presupposition satisfaction that would save the classical approach in the face of the problem of informative presuppositions, minimizing the necessity for accommodation. In the other camp, Abbott (2000) has used examples like (7) to argue that possessive definites are not presuppositional; Simons (2003) has argued much the same for factives like (6); and Gauker (1998, 2008) has used examples involving factives and definites to argue against the classical approach altogether, rejecting as inadequate the Stalnaker/von Fintel strategy for dealing with the problem.

But while factive verbs and possessive descriptions have often been taken to be canonical presupposition triggers, there is good reason to question whether that is the case. There is a long thread in

the literature, beginning with Wilson (1975) and Boër & Lycan (1976), which argues against the assumption that factives conventionally trigger presuppositions. Recent work on projection by Simons *et al.* (2010) argues that, while it is true that these triggers all tend to yield projective implications because they impose a felicity condition on the context of utterance, the classical theory assumes the wrong felicity condition. What characterizes projective meaning triggers as a class, and the factives and possessives in particular, is not that their felicitous use requires that the relevant proposition be contextually given *prior* to utterance – that is, entailed by the interlocutors' common ground, but instead that what is presupposed be *not at issue* relative to the question under discussion at the time of utterance.

What is crucial for our purposes is that there is clear empirical evidence that native speakers do *not* take factives and possessives to be presuppositional in the way assumed in the classical theory, placing a givenness requirement on prior context. Tonhauser *et al.* (2012) report detailed, methodologically controlled cross-linguistic fieldwork which argues for a distinction between two classes of projection triggers: those which impose a givenness requirement on prior context – especially the anaphoric triggers (e.g. pronouns and adverbials like English *too*) – and those which tend to trigger projection without necessarily imposing givenness – including the factives and possessive NPs. The robustness of this distinction has been corroborated experimentally by several other researchers, both for English and in other languages (Cummins, Amaral and Katsos 2013; Smith and Hall 2011; Xue and Onea 2011). Hence, there is now a growing body of evidence that a givenness requirement is not a general property defining the class of projective triggers, those that display the behavior we saw in (3) above. So although factives like *regret* and possessive definites do regularly trigger projection, from this it does not follow that they are presupposed *in the sense of being entailed by prior context*. This then argues that informative factives and possessives do not require accommodation for their felicity.

Simons *et al.* (2010) also note that there are several other kinds of expressions with projective content (that which tends to project) which clearly do not impose a givenness requirement. These include the conventional implicature triggers of Potts (2005), illustrated with the non-restrictive relative clause in (8), where the non-restrictive relative clause projects in the test contexts. (We can tell that it is non-restrictive because proper names may not be modified by restrictive relative clauses.)

(8) Georgina, who hails from Alabama, won a Pulitzer this year.
It's not the case that Georgina, who hails from Alabama, won a Pulitzer this year.
Did Georgina, who hails from Alabama, win a Pulitzer this year?
If Georgina, who hails from Alabama, won a Pulitzer this year, that proves they don't discriminate against southerners.

But such clauses in fact have an anti-givenness condition. Any of the utterances in (8) would be infelicitous in a context in which someone has just asserted that Georgina hails from Alabama. Of course, one might say that just as in the performative utterances (1) and (2) with which we began, (8) *does* involve accommodation. But this is not *presupposition* accommodation, since in neither case is what is accommodated taken to be true *prior to the utterance*.

Anaphoric expressions, like pronouns, constitute the clearest cases of presupposition triggers, and they are generally the least amenable to accommodation. Consider the opening sentence of Tom Wolfe's *The Right Stuff*, cited in this connection in Thomason (1990, 332):

(9) Within five minutes or ten minutes, no more than that, three of the others had called her on the telephone to ask her if she had heard that something had happened out there.

We do not know who these people are, or where *out there* might be. Nonetheless, in the context of the novel, where this ignorance is used as part of a particular stylistic ploy, we are willing to be patient and accommodate that there are some relevant people – one (a woman) on the phone, others in the same group of some sort – biding our time to discover more about them. But that's as far as we can get. That is, accommodation here does not involve satisfying what is presupposed by adding information to the context about the identities of the women referred to – something one cannot yet reasonably do – but merely a willingness to wait and see.

What the speaker intends as the denotation of an anaphoric expression like a pronoun or *too* must be retrieved from the context of utterance, via a prior linguistic antecedent or other highly salient information in the interlocutors' common ground. If there is no evident antecedent – as at the beginning of a novel – or there are several possible antecedents with no obvious way to exclude all but one as the speaker's intended antecedent, then accommodation is not usually possible because the descriptive content of the trigger is so impoverished: Pronouns and particles like *too* don't themselves give us many clues about who or what is presupposed. About these, von Fintel notes (2008, 154):

> [T]here cannot be accommodation with presuppositions that do not just target what is in the common ground but concern facts in the world that no manner of mental adjustment can bring into being. A particular case of that is the actual history of the conversation (the conversational record), as Beaver and Zeevat [2007] suggest. Consider Kripke's famous example (Kripke 2009[1990]):
>
> **[(10)]** Tonight, John is having dinner in New York, too.
>
> In a context in which nobody else is salient who is having dinner in New York tonight, [(10)] is unacceptable and accommodation cannot come to the rescue. Why not? Whether or not the conversation has made someone salient who is having dinner in New York tonight is part of the common ground. If the conversation hasn't made such a person salient, then it is common ground that there is no such person. And so, accommodation cannot help.

That is not to say that accommodation is never possible when anaphoric presupposition is triggered by a pronoun with no overt antecedent NP:

(11) Jacqueline's getting married. He's a soccer coach.

In (11), it seems that the retrieval of the intended antecedent for *he* proceeds via practical reasoning. *Jacqueline* is a woman's name. Until recently, if a woman was getting married, one could expect that there was a male fiancé. If interlocutors are talking about the impending marriage, in the absence of explicit mention of other men, the most relevant male would arguably be that fiancé. And resolving *he* in this way would also make the second sentence relevant to the first. Hence, reasoning to the best explanation, an addressee can abductively infer that the antecedent must be the implicitly assumed husband-to-be, and accommodate accordingly. The difference between (11) and Kripke's (10) is that in (10) we have no clues with which to infer the intended antecedent event. Resolving anaphora in (11) involves *bridging* via Jacqueline's impending marriage,[2] whereas in (10) there is nothing to permit such a bridge.

The following illustrates how accommodation may play a role in the derivation of an implication triggered solely by prosodic prominence (linguists' *focus*):

(12) (No prior discourse on a related subject)
Cecil: When are you going to China?
Lucian: Well, I'm going to [CHINA]$_B$ in [APRIL]$_A$. (Roberts 1996a)

Lucian answers Cecil's question, but his reply does more than that. He uses a prosodic focal structure with heavy emphasis both on *China* and *April*. (The annotations *B* and *A* indicate particular tonal contours typically used by native English speakers.) This implicates that Lucian is not only going to China, but to some other location as well. Here is roughly how Roberts (2012 [1996a]) and Büring (2003) take this to work: Rooth (1992) has convincingly argued that prosodic focus triggers an anaphoric presupposition – it presumes some relevant alternative set of the appropriate type. The classic literature on the semantics of questions (Hamblin 1973; Karttunen 1977; Groenendijk & Stokhof 1984) takes them to denote a set of propositions, roughly, those which are possible answers to the question. Roberts argues that the alternative set presupposed for a clause is a question under discussion derived via abstraction on the focused constituents. The prosodic contour in (12) thus conventionally presupposes that the question of interest isn't really the one Cecil asked, which would be answered merely with emphasis on *April*, but a more general one: *When are you going where?* This, in turn, implicates that the answer to Cecil's question doesn't yield a complete answer to the more general question. Otherwise, why insist on addressing the more general question? This then implicates that Lucian is going somewhere other than China; and a cooperative interlocutor, accommodating the presupposed general question and resultant implication, and taking this to be something the speaker wants to discuss, might accommodate her by asking:

(13) Oh? Where else are you going?

Practical reasoning is involved in another very common kind of case involving accommodation, the implicit contextual restriction of the domain of an operator – a quantificational determiner (*every, few, no,* . . .), adverb of quantification (*always, usually, rarely,* . . .) or modal (*would, could, should, possibly,* . . .). It has been argued that such domain restriction is presuppositional (Roberts 1989, 1995), perhaps anaphoric (von Fintel 1994; Stanley and Szabo 2000), in the latter case triggered by an implicit variable of the appropriate type at Logical Form. Domain restriction is like pronominal anaphora in that unless the addressee can retrieve the intended restriction, s/he cannot grasp the speaker's intended truth conditional meaning for the utterance. Consider:

(14) The birds will get hungry (this winter).
(15) If Edna forgets to fill the birdfeeder, she will feel very bad.
The birds will get hungry. (Roberts 1989)

Unlike (14), where (without additional prior context) the speaker is committed to a prediction that the birds will get hungry, in (15) the prediction is understood to be conditional on Edna forgetting to fill the feeder. In order to understand how the second utterance of (15) is relevant to the first, the addressee must take the modal operator which is part of the meaning of *will* to only range over worlds in which the antecedent of the preceding conditional comes to pass, so that Edna forgets.

In all the above, the reasoning involved is largely abductive: reasoning to the best explanation for the speaker's behavior, for what she said in that context. Abduction has been argued to be central to the pragmatic enrichment of utterance in general (Thomason 1990; Hobbs *et al.* 1993; Stone and Thomason 2002). Such reasoning is *holistic* in that it aims to provide a "unifying explanation" for all the puzzles involved in recognizing the speaker's intended meaning (Thomason *et al.* 2006): See (12) and (15) above, and consider (16), in which the presuppositions in question are recognized partly on the basis of an assumption of relevance of the conditional to the first sentence:

(16) Stefanie often eats out when she's travelling on business. If she doesn't know the city well, where is usually influenced by the latest Zagat ratings.

The presupposition triggered by *where* is satisfied by restricting the domain of *usually*. We bridge *the city* in the *if*-clause to some arbitrary instance of the business trips made relevant by the first utterance – reasoning that most such trips are to cities and hence taking the city in question to be Stefanie's destination away from home on an arbitrary instance of those trips. Accordingly we restrict *usually* to those occasions on which she makes such a trip. Since we have just been informed that one thing Stefanie does on such trips is eat out, and since one always eats out at some specific location, the ellipsis following *where* ("sluice") is resolved to *where she eats out*. So the interpretation must simultaneously resolve anaphora, domain restriction, and sluicing, each independently but conventionally triggered, all resolved via reasoning on the basis of the content of the conditional clauses with respect to the previous context – where the relevant content itself was not asserted but only hypothetically entertained under the scope of *often*. One might add, as well, the resolution of the Reference Time for *influence*, if one takes those to be generally assumed. More elaborate examples can be constructed. See Roberts (1995, 1996b).

Summarizing: Unlike what is often assumed in the literature, many projective triggers, including factives and possessive descriptions, are not *pre*suppositional but merely *suppositional*, backgrounding the content in question so that it is "not at issue" – merely explicitly assumed by the speaker without discussion. On the (very different) accounts of Potts (2005) and Simons *et al*. (2011), at-issue operators like negation, interrogation, etc., fail to interact with such not-at-issue content, which therefore projects – fails to be within the operators' scope. Crucially, in those cases the projective trigger together with its complement themselves identify the implication that projects. Hence, the occurrence of such triggers in contexts in which the projective implications are informative poses no problem for the recognition of what is assumed by the speaker, and hence does not call for accommodation *in the sense of repair of a faulty context of utterance*. If there is accommodation in such cases, it is the willingness of the addressees to accept the supposition as uncontroversial, adding it to the Common Ground.

In contrast, accommodation is not usually available for anaphoric presuppositions like those triggered by pronouns or *too*, and by ellipses. This is because the descriptive content of a typical anaphoric trigger is not usually sufficiently rich to identify exactly what is presupposed – the intended antecedent, etc. Hence, their use presupposes salience in the actual discourse, which typically cannot be accommodated; as von Fintel (2008) notes, such salience is a fact about the world, which either does or does not obtain. Cases where anaphoric presuppositions *do* appear to be accommodable involve bridging on the basis of practical (abductive) reasoning, which identifies what is most likely presupposed *given the context of utterance*.

The problem of presupposition recognition is all the more challenging when there is no overt trigger, as is typical in domain restriction. Arguably, as we will discuss in section 22.4, the recognition in such cases *that* the speaker presupposes a restricted domain arises via assumptions about the relevance of the utterance to prior context.

Hence, it seems that much of the debate over accommodation is mis-aimed. The central question about accommodation is not whether it can be used to save the theory of presupposition in the face of examples with informative presuppositions, but how addressees come to recognize exactly what a speaker presupposes in those cases where what is to be accommodated is merely implicitly triggered, or is explicitly triggered but with the presupposed content underdetermined by the trigger and no overt, sufficiently salient antecedent.

22.3 The Character of Accommodation

What is it to accommodate, and why do we do it? Though Lewis introduces *accommodation* as a term of art, it was surely chosen with a view to its ordinary meaning, to wit:

accommodate:
1. to make fit, suitable, or congruous
2. to bring into agreement or concord: reconcile (Merriam-Webster Dictionary online)

An extension of this sense noted by several dictionaries is something like 'to oblige', as in *the hostess was willing to accommodate her friend*. Near-synonyms for *accommodation* are given by the same dictionary as *adaptation, adjustment*.

All of these senses seem relevant to the technical notion of accommodation of interest to Lewis. Here is the characterization due to Thomason (1990, 343–4):

> . . . [A]ccommodation is a special case of obstacle elimination, a form of reasoning discussed in Allen 1983. Obstacle elimination consists in (1) recognizing the plan of your interlocutor; (2) detecting obstacles to the plan in the form of certain false preconditions of subgoals belonging to the plan; (3) adopting the goal of making these preconditions true; (4) forming a plan to carry this out; and (5) acting on this plan. Step (1) is plan recognition. Step (3) is cooperative goal adoption. The other steps employ forms of reasoning that figure in noncooperative planning by isolated agents. . . .
>
> The principle behind accommodation, then, is this:
>
> > Adjust the conversational record to eliminate obstacles to the detected plans of your interlocutor.
>
> If the term hadn't already been claimed, this could well have been called the *cooperative principle*.

So accommodation takes place only in circumstances where one agent, recognizing the plans, goals, and intentions of another, can adjust her plans so as to help further the other's goals.

It is not for nothing that Thomason invokes Grice. According to Grice (1957), meaning (his *meaning*$_{nn}$) involves a complex set of intentions on the part of the speaker. In sympathy with that characterization, we might informally characterize an utterance as linguistic behavior which implements a plan on the part of the speaker S to get her addressee A to recognize a certain content (propositional, interrogative, etc.), which is then to be considered by A for possible addition to their shared information. Since the conventional content of an utterance typically underdetermines its intended meaning, a cooperative addressee accommodates S's plan by filling in the blanks – adding content which is not explicit in the utterance as a function of its conventional form but which can be readily inferred to be part of S's intended meaning in so uttering, that is, to be part of her plan. In this way, A – even if he ultimately rejects the intended content (as untrue, inappropriate, etc.) – cooperates with S in realizing her plan to *convey* that meaning. The first step is plan recognition – A must retrieve the intended extra-conventional content, as illustrated in section 22.2, partly as a function of grasping 'what S is getting at'. Should the plan involve goals which have unsatisfied preconditions, that is potentially problematic for S's realization of her plan. If it is within A's power to do so, A cooperatively rectifies this situation:

> Acting as if we don't have a flat tire won't repair the flat; acting as if we know the way to our destination won't get us there. Unless we believe in magic, the inanimate world is not accommodating. But *people* can be accommodating, and in fact there are many social situations in which the best way to get what we want is to act as if we already had it . . . (Thomason 1990, 342)

Suppose the obstacle to conveying the intended meaning of an utterance is a problem with presupposition satisfaction: a pronoun fails to have an obvious antecedent. If bridging is reasonable in the context, as in (11), (12) or (16), A can accommodate S's plan by behaving as if the presupposition

of an antecedent were satisfied. So long as the resulting interpretation is relevant and coherent with the rest of the text, the result is as good as if there were an overt antecedent to begin with, so there is no impediment to accommodation.

Thomason focuses on another type of accommodation, not involving presupposition, where in order to grasp the speaker's intended meaning a conversational implicature must be accommodated (1990, 354):

(17) [husband to wife, preparing to leave in the morning:] I didn't tell you that I'll need the car this afternoon.

By his utterance, the husband in (17) manages to convey to his wife that he does need the car that afternoon – one might reasonably say that he *means* (in Grice's sense) that he needs the car. But how? Thomason points out that the conventional content of the utterance itself is trivial and obvious to both parties to the conversation. Given only this, one might ask whether the husband intends anything more than one's barber means by beginning a conversation with the trivial *Nice weather we're having*. But there is more to a conversation, any conversation, than that:

> In almost any sort of conversation we always feel compelled to reconstruct the plans of our interlocutors; and we want to see how the message we ascribe to them fits into a model of their purposes. If we can't do this, we are likely to resort to accommodation in order to make it fit. Triviality is one feature that makes it more difficult to achieve this sort of fit; and this is why an assertion that is literally trivial is likely to give rise to implicatures . . . The success of the discourse strategy [in (17)] . . . depends in part on the fact that the sentence is negative and comes at the beginning of a discourse unit. Such sentences invite the hypothesis that they express *a lack*, and in fact point to an obstacle in a plan (of the speaker's, or perhaps someone else) . . . For instance, if I begin a conversational unit by saying "There isn't a doorstop in this room," it would be appropriate for you to say "Why would you want a doorstop?" and it would be disingenuous of me to say "I didn't say I wanted a doorstop." Thus, I can suppose that saying "I didn't tell you I'll need the car this afternoon" will launch a search on my wife's part for a plan of mine that would be thwarted by my not telling her that I needed the car. If she has a normal ability to recognize domain plans [those pertaining to objectives in the world], this should suffice to meet my discourse goal. (Thomason 1990, 353–4)

Hence, on the assumption that the husband's utterance of (17) plays a role in a larger plan – that he's getting at something of interest, the wife abductively infers that his plan requires the use of the car that afternoon, and accommodates this implication. Thomason takes the accommodated content to be a *conversational implicature*.

Though (17) doesn't involve presupposition, it is of interest here because, as in the cases of domain restriction noted in section 22.2, there is no overt trigger for the implication in question, and the question arises of how we recognize what the speaker means in that respect. It seems that the same kind of practical, abductive reasoning is involved as was observed in the domain restriction cases, arguing that there is some common denominator between the accommodation of implicatures, and presupposition recognition and accommodation.

Consider again the examples discussed in section 22.2. We saw three types of cases that might be argued to involve a cooperative adjustment of the context on the part of the addressee:

1. cases in which some explicit content is taken for granted without being asserted, but not presupposed – backgrounded implications like those triggered by the factives and possessive NPs;
2. cases involving true anaphoric presuppositions where a presuppositional trigger puts a conventional constraint on the kind of context in which the expression in question is felicitous and interpretable but that constraint is not satisfied in the context of utterance; and

3. cases without an overt trigger in which some additional content must be abductively inferred in order to make sense of why (or how) the speaker is saying what he's saying – implicit domain restriction and conversational implicatures.

Only in the first, backgrounded, cases do we have explicit expression of what is to be accommodated. This type of accommodation doesn't involve repair, only cooperative acceptance of that content. The second and third types do involve repair. First, there is a recognition that something is "missing" (conventionally triggered in the first case, conversational in the third). Then practical reasoning based on abduction is used to (try to) ascertain what that missing content might be. This is Thomason's obstacle identification and elimination: detecting the speaker's plan (her meaning) and obstacles to that plan in the form of certain false preconditions of subgoals belonging to the plan (what's presupposed). And finally, *ceteris paribus*, willingness to act as if the missing implication were true: accommodation to remove the obstacles.

But, in the third class, how do we recognize that these obstacles exist? And, in both the non-explicit classes, how do we retrieve what is to be accommodated to remove those obstacles?

22.4 The Role of the Scoreboard in Accommodation

Lewis's view of the scoreboard is in keeping with the approach to philosophical analysis once disparagingly labeled the *Canberra Plan* (O'Leary-Hawthorne and Price 1996, 291, n.23): Its content "depends on the history of the conversation in the way that score should according to the rules" (1979, 346). I take it that this is a throwback to Lewis's excellent advice in "General Semantics" (1970, 23):

> In order to say what meaning *is*, we may first ask what a meaning *does*, and then find something that does that.

The scoreboard in Lewis (1979) seems to tell us something about what Lewis takes a context of utterance to be, paraphrasing Lewis in the earlier paper:

> In order to say what a context of utterance *is*, figure out what it *does* – how it interacts with the rules of the language game to yield attested interpretations – and then find something that does that.

The nature and role of accommodation is best appreciated in light of this general project. We begin by exploring what Lewis might have intended by using the scoreboard metaphor, considering in particular the function and "kinematics" of the type of scoreboard he uses as a concrete comparison: the scoreboard in a baseball game. My characterization may go beyond what Lewis himself envisioned, certainly beyond what he says in this paper. But that is perhaps to be expected. Insightful proposals often have broader implications than are initially recognized by the author.

The baseball scoreboard tracks several kinds of information, which change in different ways over the course of play. One bit tracks the length of a game, how many innings have elapsed; under most circumstances, after nine innings the game is over, each inning affording a turn to each of the two teams. But there are two other kinds of information tracked on the board. One is the current score in the narrow sense, pertaining to the essential goal of the game: the number of runs which each team has earned. At the end of play, the team with the highest score in this sense wins. This score is monotonic: once a run is scored, it stays on the board for the duration of the game. The other type of information is useful for keeping track of which team is entitled to make which moves at that

juncture in play. This second type of information is non-monotonic, reset at certain points according to the rules. It includes which team is at-bat; the number of players who have been struck out during that turn (after three outs the team loses its turn); and the number of balls and strikes during an at-bat by a single would-be hitter (the balls counting against the pitcher, the strikes against the hitter; three strikes and he's out).[3]

Similarly (though Lewis doesn't explicitly explore this aspect of the analogy), the language scoreboard contains two kinds of information. The first, the Common Ground ("sets of presupposed propositions"), pertains to what is arguably the point of the language game. Stalnaker (1979) takes the goal of discourse to be to discover what the world is really like, and models this with the Context Set; this is the set of worlds compatible with the interlocutors' Common Ground. The goal is to figure out which world they are really in, ultimately reducing the Context Set to the unique actual world. Ideally the addition of information to the Common Ground is monotonic (ignoring the necessity for belief revision, a type of repair), and correspondingly, so is the reduction of the Context Set. But there are other types of information which play a crucial role in interpretation but are non-monotonic, tracking the state of play itself. For example, salience shifts from move to move in the language game, a function of what's most immediately under discussion. Salience doesn't count in the final score, but like the number of strikes a batter has, it plays a crucial role in constraining play, here the interpretive process, notably for anaphora resolution.

Similarly, recent work (Ginzburg 1994, 2012; Roberts 1996a, 2004) has argued that the (set of) question(s) under discussion (QUD) at a given point in discourse plays a central role in interpretation, in phenomena ranging from prosodic focus (Roberts 1996a; Büring 2003), domain restriction (Roberts 1996b; Beaver and Clark 2008), ellipsis resolution (Anderbois 2011; Ginzburg 2012), rhetorical relations (Jasinskaja 2007), and implicature generation (van Kuppevelt 1995; van Rooij 2003) to presupposition recognition and projection (Simons *et al.* 2010) and the determination of salience (Roberts 2011b). To the extent that interpretation is influenced by the QUD across such a broad range of prima facie unrelated phenomena, this argues that it is a central part of what interlocutors track about the discourse in which they are participating.

Roberts's (1996a, 2004) characterization makes this role intuitively natural for an approach like Lewis' game theoretical view of language by taking the immediate QUD to establish the immediate subgoal (answering the QUD) of the overarching discourse goal, answering the Big Question: *What is the way things are?* to enable the reduction of the Context Set. Questions propose new discourse goals for cooperative adoption. Imperatives are like questions in posing goals to the interlocutors – but these are domain goals, proposals for non-linguistic action. There are logical constraints on the relations between various questions under discussion and related domain plans of the interlocutors, so that not just any proposed QUD is felicitous in a given discourse context. Utterances, *qua* moves in the game, are felicitous to the extent that they are relevant to the adopted QUD. An assertion should offer a partial answer to the QUD; a question should suggest a subquestion of the QUD; a suggestion or order should propose an action which would further answering the QUD. If a group of collaborating agents are rational, the goals they adopt, and the plans they form to achieve them, are consistent, and, in particular, discourse goals (QUD) should subserve overarching domain goals.

Adopting a goal involves adopting an intention. Thus, collectively the questions under discussion and the domain plans they subserve place an intentional structure on rational discourse interaction. Participants engaged in this collaborative game cooperate by attempting to implement these plans, most immediately by addressing the QUD. So like balls and strikes, QUDs and associated domain plans and intentions are only temporarily on the scoreboard, removed once they have been achieved, at which point a new goal may be posted (or the game – the discourse – is over).

This conception of the scoreboard is consistent with what Lewis has to say about his "middle way" of modeling context. He tells us that on this conception (with my comparisons in square brackets):

- "[T]he components of a conversational score at a given stage are abstract entities. They may not be numbers, but they are other set-theoretic constructs" (1979, 345). The score, as in baseball, is a tuple of such abstract entities. Hence, rules for update (the "kinematics of score") are not themselves part of this abstract score so-defined, but "enter only in a roundabout way into the definition of score."
- The elements of the conversational score may include (among other things):
  - sets of presupposed propositions (346) [cf. the Common Ground in Stalnaker 1979; Clark 1996]
  - the interlocutors' plans (357)
  - comparative salience of the entities under discussion (349).
- The rules for updating context ("the kinematics of score") "underdetermine the evolution of score" (346), and hence the score includes non-linguistic, non-conventional content. This is implicit in the kinds of elements just listed. Sets of presupposed propositions are generally assumed to be those which, like Stalnaker's (1979) Common Ground (CG), the interlocutors all (purport to) take to be true for whatever reason, hence not only those propositions proffered in the current conversation. Certainly we do not usually directly discuss what is salient.
- "[I]t is possible that score sometimes evolves in a way that violates the rules." We may flout maxims, and even utter something denoting P when we mean not-P (sarcasm, irony), etc.

Including interlocutors' plans – especially in the guise of questions under discussion – on the scoreboard permits us to capture how those plans both drive and constrain the behavior of the participants in the game. In particular, it offers insight into the role of the scoreboard in presupposition recognition, identifying the obstacles to the inferred plan of the speaker. As we saw above, this depends on the ability to abductively infer what the speaker meant: to reason about the best explanation for her speech act, given its conventional content, in light of *what we already know of her plans and intentions*. This is key to abduction, as utilized extensively in Planning Theory in Artificial Intelligence (see Allen and Perrault 1980; Appelt 1985; Thomason and Hobbs 1997; Stone 2004). Thomason, Stone and DeVault (2006) apply this approach to extend Thomason's (1990) characterization of accommodation. For them (with my comments and glosses in square brackets):

- Meaning is a kind of complex intention. [Grice 1957]
- An intention is a kind of complex information state, including:
  - a goal ['intend to…']
  - a plan ['by…']
  - preconditions ['what must be the case in order for the plan to succeed']
- Interpretation, or meaning-recognition, is plan-recognition. [Planning Theory]
- The preconditions in a linguistic meaning are the presuppositions of the relevant utterance.
- So presupposition recognition (including anaphora resolution) is a species of plan recognition: Recognizing or retrieving the preconditions on the meaningfulness or truth of the utterance in question.
- Accommodation itself is the "removal of obstacles"…

The clearer one's interlocutors' plans and goals, the easier it is to bootstrap this process of obstacle identification and removal. Taking the resolution of the QUD as the immediate cooperative goal of the interlocutors, one which defines what it is to be relevant, and hence felicitous, at that point in play, thus plays a central role in recognizing the probable plan of the speaker in making that particular utterance at that particular point in the game, the meaning he aimed to convey. If the speaker is

competent (understands the rules of the game, including the requirement of relevance to the QUD) and cooperative, then whatever she said was intended to address the QUD. Then one should resolve anaphora, restrict quantificational domains, enrich with implicatures, etc., as necessary and reasonable in view of what's salient in such a way as to take her utterance maximally relevant to that question. Salience itself can be argued to be a function of the QUD. This gives addressees a lot of leverage in presupposition recognition, especially in the cases requiring accommodation where what is presupposed, and even *that* it is presupposed, is inexplicit.

Accommodation is natural to the extent that the meaning to be accommodated is the uniquely obvious abductively inferred means of making the utterance relevant to the interlocutors' goals at that point in play. *Accommodation in such a case is easy and natural*, not something the average interpreter even notices.

The theory of the scoreboard sketched here is a causally efficacious content theory of context, one in which the scoreboard contains not only the score in the narrow sense (the Context Set), but also that temporary information which reflects the way in which the scoreboard controls the language game, playing a role in the adoption of directives (queries posed by questions, suggestions by imperatives), in coordination (helping to track shared goals and background information), and in useful adaptation (accommodation). Through making clear the interlocutors' common goals, the scoreboard thus both facilitates and constrains accommodation.

22.5 Limits on Accommodation

> It is not good conversational practice to rely too heavily on rules of accommodation . . . but] Confusing shifts of salience and reference are not as bad as falsity, trivial truth, or unwarranted assertion. (Lewis 1979, 350)

It has often been claimed that accommodation is too powerful. Can we offer a predictive theory of accommodation, one in which it is appropriately constrained and in particular accounts for presupposition *failure*? Yes, we can, in light of the characterization of the context of utterance sketched in the previous section. Quite simply, accommodation is constrained by requirements of consistency and coherence, these gauged with respect to the scoreboard.

We cannot always accommodate a speaker whose meaning fails to be clear, even if the failure to accommodate leaves the resulting utterance infelicitous or uninterpretable. In Kripke's example (10) repeated here, *too* is anaphoric, but we simply cannot retrieve an antecedent because the example is given more or less out of the blue:

(10) Tonight, John is having dinner in New York, too.

Note that, like the pronouns and ellipses in (11), (12), and (16), *too* can sometimes take an antecedent which is only implicitly available, as in (18):

(18) [Two women are standing at a bus stop on a rainy day. A car drives through a puddle along with curb, splashing one of the women with muddy water. The second woman turns to her and says:] A car splashed ME this morning, too.

In (18), there *is* a relevant antecedent for *too*, albeit non-linguistically introduced: the proposition that a car just splashed the addressee, presumably an event still sufficiently distressing to be foremost in her mind. So it's not even clear that (18) involves accommodation. But in (10) there clearly *is* no

antecedent for *too* that's relevant in the context of utterance. We have no clue about what the speaker may mean (in Grice's sense) in using *too* in his utterance. In keeping with the observations of Thomason (1990, 342) and von Fintel (2008, 154) above, no amount of accommodation here can make true a presupposition (that there is a salient antecedent) which is manifestly false. Therefore, in such cases accommodation is not acceptable.

Another type of example involves *only* (see Roberts 2011a for extended discussion). When we have an utterance of the form *only SUBJECT VPs*, as in (19), we call the result of removing *only* from that clause the *prejacent*:

(19) Only Lucy came to the party.
prejacent: Lucy came to the party.

Only is a trigger for projection; that is, its prejacent tends to project in the test contexts for projection, as we see in (20):

(20) Projection tests for the prejacent of *only*:
It's not the case that only Lucy came to the party./Not only Lucy came to the party.
Did only Lucy come to the party?
If only Lucy came to the party, it must have been pretty quiet.
Maybe only Lucy came to the party.

All of the variants in (20) seem to implicate that Lucy came to the party. Hence, (21) (with a single speaker) sounds like a contradiction (marked as infelicitous: #):

(21) Did only Lucy come to the party? #Of course, SHE didn't.

But just like a possessive NP or the complement of a factive verb, the prejacent of *only* is pretty clearly not *pre*supposed, but only *supposed*, or backgrounded, so that the prejacent can be novel in the Common Ground. One argument for this is the felicity of B's response in (22), where the truth of the prejacent is the very question under discussion:

(22) A: Did Lucy come to the party?
B: Actually only Lucy came.

Then just like the outcome of a performative like (1) or (2), or after utterance of novel factive complements like (6) or (7), an addressee who finds the prejacent of (19) or (20) or that of B's answer in (22) uncontroversial will be inclined to accommodate it.

But not always. When the prejacent is controversial, or, especially, would contradict information in prior context, it fails to project, that is, is not accommodated to the Common Ground:

(23) And contrary to what many say I found the level of violence high but not excessive. This isn't only a "shoot 'em up" pointless movie; there's more than just stage blood. (web example reported in Beaver and Clark 2008, 235)

(24) [about a family where women generally have lots of kids]
QUD: How many kids does each of these siblings have?
Mary's the blacksheep. As far as I know she doesn't have any kids, but I can't remember for sure. Maybe she only has one kid? George, do you remember?

In (23), the writer is clearly arguing that the movie in question is worthwhile, hence it would be inconsistent for the prejacent of the second clause 'this is a shoot 'em up pointless movie' to project from under negation. In (24) the question of how many kids Mary has is under discussion, and the speaker makes it clear that for all she knows Mary may not have any; hence the prejacent 'she has one kid' doesn't project from under the modal in the third sentence. To sharpen our grasp of the flexibility of the projective behavior of the prejacent, see how it contrasts with that of the non-restrictive relative clause *who has one kid* in (25), which obligatorily projects, leading to a contradiction with the last sentence:

(25) QUD: How many kids does Mary have?
George told me that Mary, who has one kid, is the blacksheep of her family. She doesn't have any kids.

These examples illustrate the requirement of *consistency*: The interlocutors' Common Ground must be logically consistent. Nothing can be accommodated which would make it inconsistent.

Another kind of case where accommodation fails involves prosodic focus on the wrong word(s) in an utterance. In answering the question in (26), focus should fall on the object *the pasta*; while after a different question in (27), for an answer with the same propositional content as that in (26) to be felicitous, it should have focus on the transitive verb *ate*. In each case, focus should fall on that part of the answer which is informative with respect to the preceding question, correlating with the *wh*-word in the question:

(26) What did Karen eat?
A1: She ate the PASTA.
A2: #She ATE the pasta.
(27) What did Karen do with the pasta?
A1: #She ate the PASTA.
A2: She ATE the pasta.

Note that focus in those examples doesn't mark what's new information in the sense of content not previously mentioned: In (28), both *praise* and the direct object *him* are already salient from the question. But here, too, it must be the direct object, correlating in its grammatical role with that of *who* in the question, which receives focus:

(28) Who did John_i's mother praise?
A1: She praised HIM_i.
A2: #She PRAISED him_i.

However, some felicitous answers are prima facie non-congruent with the preceding question in this respect, as we saw from (12), repeated here:

(12) (No prior discourse on a related subject)
Cecil: When are you going to China?
Lucian: Well, I'm going to $[\text{CHINA}]_B$ in $[\text{APRIL}]_A$.

The difference, I have argued (Roberts 1996a), is that in (12) the addressee can accommodate the question presupposed by the prosodic structure of Lucian's answer: 'when is Lucian going where?',

taking both the accommodated question and Cecil's actual question to participate in a larger strategy of inquiry. But this is only possible because the resulting strategy is felicitous and coherent: The actual question is a subquestion of the accommodated question, hence poses a consistent subgoal. In (26), by contrast, there is no obvious way to accommodate the question 'what did Karen do with the pasta', taking it to be part of a strategy of inquiry with the preceding overt question. Accordingly, the discourse fails to be coherent, and accommodation fails.

Roberts (2004) argues that the central feature of a coherent discourse is that it reflects a rational strategy of inquiry with respect to the questions under discussion. Various kinds of rhetorical strategies which also bear on cohesion (Kehler 2002; Asher and Lascarides 2003) can then be characterized as types of strategies of inquiry (see also the D-trees of Büring 2003), hence themselves a reflex of relevance to the topical structure of the discourse (Kehler 2009). Other features that have been argued (e.g. in Halliday and Hasan 1976) to play a role in cohesion – anaphora (including temporal anaphora), ellipsis, conjunction – can be argued to hinge on relevance, as defined in terms of addressing the QUD and broader strategy of inquiry.

This brief consideration of a few examples is intended to suggest that the rules that constrain accommodation are not particular to that function. Consistency and cohesion are general rules governing what constitutes a well-formed discourse. Both these requirements, in turn, arguably follow from the overarching goal of the language game and its intentional structure. If the goal is to figure out what world we're in, inconsistency, leading to an empty Context Set, is inconsistent with those goals. Then a rational agent who is committed to that goal will seek to avoid inconsistency. Similarly, if one intends to observe the rules of the game, in view of how they support the goal of sharing information, one avoids non sequiturs, since they yield incoherence, and incoherence makes it difficult to retrieve one's meaning.

22.6 Conclusion

In this chapter I have aimed for:

- exegesis: What did Lewis mean, and how does it fit in his overall program for understanding linguistic meaning? I have argued that his proposals in "Scorekeeping…" (1979) are best understood in light of his larger program of illuminating the nature of linguistic convention and communication. In "Scorekeeping…" Lewis sketches a way of understanding how his earlier characterization of the language game can be extended to permit insight into complex, non-indexical context-sensitivity.
- explication: What is accommodation, especially presupposition accommodation, and what role does it play in interpretation? I have argued that presupposition accommodation, a kind of repair, is perhaps less common and more restricted than is sometimes assumed, but that the broader phenomenon, including *supposition*-accommodation, is far *more* common than is generally recognized. In keeping with Thomason (1990), accommodation to remove obstacles to the speaker's evident semantic goals is a natural part of the language game, of recognizing a speaker's intended meaning.
- explanation: What's the scoreboard of a language game, and what kind of role might it play in explaining the resolution of context-sensitive expressions? A causally efficacious model of context as scoreboard will crucially involve tracking the immediate discourse goals of the interlocutors (the QUD). Such a model of context both drives accommodation – to permit utterances to be understood as relevant to those goals, and constrains it – to guarantee consistency and coherence, both these desiderata understood in terms of logical constraints on the elements of the score.

We began by noting that some see accommodation as too strong, making a theory which uses it unfalsifiable. But I have suggested here that an *appropriate, constrained notion of accommodation* is a *sine qua non* for progress in understanding not only presupposition, but phenomena like implicature generation, the recognition of intended lexical content in the face of vagueness and imprecise use, lexical coercion, and various types of semantic variation, including dialect variation and language change. The exploration of such matters is beyond our purview here, but I believe that in all these cases we accommodate – go along with, or accord our understanding with – the speaker's semantic intentions, so long as (a) those intentions are evident and (b) the resulting context is consistent and coherent. In fact, we cannot generally *understand* what a speaker means unless we are at least temporarily willing to be accommodating in this way, even if we ultimately reject the intended contribution to the score (e.g. as false). Accommodation so described needn't be a repair, but is instead a natural, common feature of the activity of conveying linguistic meaning in context through underspecified conventional content.

Acknowledgments

I gratefully acknowledge the support of grant 0952571 from the National Science Foundation, with co-PIs and invaluable colleagues David Beaver, Mandy Simons, and Judith Tonhauser. In addition, I had the privilege of completing this work while a Fellow at the National Humanities Center (Research Triangle, North Carolina). Without their support, and the continuous support of my home institution, The Ohio State University, none of this work would have been completed. Besides my co-PIs, I am also indebted to B. Chandrasekaran, Scott Martin, Barbara H. Partee, Kevin Scharp, Stewart Shapiro, Eric Snyder, Richmond Thomason, and audiences in the OSU Pragmatics Working Group, the Department of Philosophy at Rutgers University and the Universitat Autonoma de Barcelona, whose comments on preliminary versions of the chapter were very helpful.

Notes

1 Karttunen (1974, 191) discusses the problems that examples like his (6) raise for his theory, and effectively proposes accommodation in such cases:

> I think we can maintain that a sentence is always taken to be an increment to a context that satisfies its presuppositions. If the current conversational context does not suffice, the listener is entitled and expected to extend it as required. He must determine for himself what context he is supposed to be in on the basis of what was said and, if he is willing to go along with it, make the same tacit extension that his interlocutor appears to have made. This is one way in which we communicate indirectly, convey matters without discussing them.

2 Bridging (Clark 1974) is anaphora resolution via a pragmatically related, salient, but non-coreferent entity – here the man and marriage.

3 Thanks to Barbara H. Partee (pers. comm.) for correcting my description of baseball scoring.

References

Abbott, B. (2000). Presuppositions as nonassertions. *Journal of Pragmatics*, 32(10), 1419–37.

Abbott, B. (2008). Presuppositions and common ground. *Linguistics and Philosophy*, 31, 523–38.

Allen, J.F. (1983). Recognizing intentions from natural language utterances. In M. Brady & R. Berwick (Eds.), *Computational models of discourse*, pp. 107–66. Cambridge, MA: MIT Press.

Allen, J.F. & Perrault, C.R. (1980). Analyzing intention in utterances. *Artificial Intelligence*, 15, 143–78.
Anderbois, S. (2011). Sluicing as anaphora to issues. *Proceedings of semantics and linguistic theory 20*. Ithaca, NY: CLC Publications.
Appelt, D. (1985). Planning English referring expressions. *Artificial Intelligence*, 26(1), 1–33. Reprinted in B. Grosz, K. Sparck Jones & B. Webber (Eds.) (1986), *Readings in natural language processing*. San Francisco: Morgan-Kaufmann.
Asher, N. & Lascarides, A. (2003). *Logics of conversation*. Cambridge: Cambridge University Press.
Beaver, D. (2001). *Presupposition and assertion in dynamic semantics*. Studies in Logic, Language and Information. Stanford, CA: CSLI Publications.
Beaver, D. & Clark, B. (2008). *Sense and sensitivity*. Oxford: Wiley-Blackwell.
Beaver, D. & Zeevat, H. (2007). Accommodation. In G. Ramchand & C. Reiss (Eds.), *The Oxford handbook of linguistic interfaces*, pp. 503–38. Oxford: Oxford University Press.
Boër, S.E. & Lycan, W.G. (1976). The myth of semantic presupposition. Mimeograph, Indiana University Linguistics Club.
Büring, D. (2003). On D-Trees, Beans, and B-Accents. *Linguistics and Philosophy*, 26 (5), 511–45.
Clark, H.H. (1974). Bridging. In R.C. Schank & B.L. Nash-Webber (Eds.), *Theoretical issues in natural language processing*. New York: Association for Computing Machinery.
Clark, H.H. (1996). *Using language*. Cambridge: Cambridge University Press.
Cummins, C., Amaral, P. & Katsos, N. (2013). Backgrounding and accommodation of presuppositions: an experimental approach. In E. Chemla, V. Homer & G. Winterstein (Eds.), *Proceedings of Sinn and Bedeutung 17*, pp. 201–18.
Gauker, C. (1998). What is a context of utterance? *Philosophical Studies*, 91(2), 149–72.
Gauker, C. (2008). Against accommodation: Heim, van der Sandt and the presupposition projection problem. *Philosophical Perspectives: Philosophy of Language*, 22(1), 171–205.
Ginzburg, J. (1994). An update semantics for dialogue. In H. Bunt (Ed.), *Proceedings of the Tilburg International Workshop on Computational Semantics*. Tilburg: ITK.
Ginzburg, J. (1995). Resolving questions, Part I and Part II. *Linguistics and Philosophy*, 18(5), 459–527 and 18(6), 567–609.
Ginzburg, J. (2012). *The interactive stance: meaning for conversation*. Oxford: Oxford University Press.
Grice, H.P. (1957). Meaning. *Philosophical Review*, 66, 377–88.
Groenendijk, J. & Stokhof, M. (1984). Studies on the semantics of questions and the pragmatics of answers. PhD dissertation, University of Amsterdam.
Halliday, M.A.K. & Hasan, R. (1976). *Cohesion in English*. London: Longman.
Hamblin, C. (1973). Questions in Montague English. *Foundations of Language*, 10, 41–53. Reprinted in B. Partee (Ed.) (1976), *Montague grammar*. Austin: University of Texas Press.
Heim, I. (1982). The semantics of definite and indefinite noun phrases, PhD dissertation, University of Massachusetts-Amherst.
Heim, I. (1983). On the projection problem for presuppositions. *Proceedings of WCCFL 2*, pp. 114–26. Stanford, CA: Stanford University Press.
Hobbs, J.R., Stickel, M.E., Appelt, D.E. & Martin, P. (1993). Interpretation as abduction. *Artificial Intelligence*, 63, 69–142.
Jasinskaja, E. (2007). Pragmatics and prosody of implicit discourse relations. PhD thesis, University of Tübingen.
Karttunen, L. (1973). Presuppositions of compound sentences. *Linguistic Inquiry*, 4:169–193.
Karttunen, L. (1974). Presupposition and linguistic context. *Theoretical Linguistics*, 1, 181–94. Reprinted in S. Davis (Ed.) (1991), *Pragmatics: a reader*, pp 406–15. Oxford: Oxford University Press.
Karttunen, L. (1977). Syntax and semantics of questions. *Linguistics and Philosophy*, 1, 3–44.
Kehler, A. (2002). *Coherence, reference, and the theory of grammar*. Stanford, CA: CSLI Press.
Kehler, A. (2009). Ellipsis and anaphora in a QUD model of discourse. Talk presented at the University of Michigan Workshop in Philosophy and Linguistics.
Kripke, S.A. (2009 [1990]). Presupposition and anaphora: remarks on the formulation of the projection problem. *Linguistic Inquiry*, 40(3), 367–86.

Lewis, D. (1969). *Convention: a philosophical study*. Cambridge, MA: Harvard University Press.
Lewis, D. (1970). General semantics. *Synthese*, 22, 18–67.
Lewis, D. (1979). Scorekeeping in a language game. In R. Baüerle, U. Egli & A. von Stechow (Eds.), *Semantics from a different point of view*. Berlin: Springer.
O'Leary-Hawthorne, J. & Price, H. (1996). How to stand up for non-cognitivists. *Australian Journal of Philosophy*, 74, 275–92.
Partee, B.H. (2004). *Compositionality in formal semantics*. Oxford: Blackwell.
Potts, C. (2005). *The logic of conventional implicatures*. Oxford: Oxford University Press.
Roberts, C. (1989). Modal subordination and pronominal anaphora in discourse. *Linguistics and Philosophy*, 12, 683–721.
Roberts, C. (1995). Domain selection in dynamic semantics. In E. Bach, E. Jelinek, A. Kratzer & B.H. Partee (Eds.), *Quantification in natural languages*, pp. 661–700. Dordrecht: Kluwer.
Roberts, C. (2012[1996a]). Information structure: toward an integrated theory of formal pragmatics. In J.H. Yoon & A. Kathol (Eds.), *OSUWPL volume 49: papers in Semantics*. Columbus, OH: OSU Department of Linguistics. Reprinted, with a new afterword, in C. Roberts (in press), *Semantics and pragmatics*.
Roberts, C. (1996b). Anaphora in intensional contexts. In S. Lappin (Ed.), *Handbook of semantics*. Oxford: Blackwell.
Roberts, C. (2004). Discourse context in dynamic interpretation. In L. Horn & G. Ward (Eds.), *Handbook of contemporary pragmatic theory*, pp. 197–220. Oxford: Blackwell.
Roberts, C. (2011a). A case study in projective meaning. In B. Partee, M. Glanzberg & J. Skilters (Eds.), *Formal semantics and pragmatics: discourse, context, and models*. The Baltic International Yearbook of Cognition, Logic and Communication. Manhattan, KS: New Prairie Press.
Roberts, C. (2011b). Solving for interpretation. Talk at the Workshop on Meaning and Understanding at the Centre for Advanced Study, Oslo, June.
Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, 1(1), 75–116.
Simons, M. (2003). Presupposition and accommodation: understanding the Stalnakerian picture. *Philosophical Studies*, 112(3), 251–78.
Simons, M., Tonhauser, J. Roberts, C. & Beaver, D. (2010). What projects and why. In N. Li & D. Lutz (Eds.), *Semantics and linguistic theory (SALT) 20*, 309–27. Ithaca, NY: CLC Publications.
Smith, E.A. & Hall, K. (2011). Projection diversity: experimental evidence. Proceedings of the ESSLLI 2011 Workshop on Projective Meaning, Ljubljana, Slovenia.
Stalnaker, R. (1973). Presuppositions. *Journal of Philosophical Logic*, 2, 447–57.
Stalnaker, R. (1979). Assertion. In P. Cole (Ed.), *Syntax and Semantics*, p. 9. New York: Academic Press.
Stalnaker, R. (1998). On the representation of context. *Journal of Logic, Language and Information*, 7, 3–19.
Stalnaker, R. (2008). A response to Abbott on presupposition and common ground. *Linguistics and Philosophy*, 31, 539–44.
Stanley, J. & Szabó, Z. (2000). On quantifier domain restriction. *Mind and Language*, 15(2–3), 219–61.
Stone, M. (2004). Intention, interpretation and the computational structure of language. *Cognitive Science*, 5, 781–809.
Stone, M. & Thomason, R. (2002). Context in abductive interpretation. In J. Bos, M.E. Foster & C. Mathesin (Eds.), *EDILOG 2002: Proceedings of the Sixth Workshop on the Semantics and Pragmatics of Dialogue*, pp. 169–76. Edinburgh: Cognitive Science Centre, University of Edinburgh.
Thomason, R. (1990). Accommodation, meaning, and implicature: interdisciplinary foundations for pragmatics. In P. Cohen, J. Morgan & M. Pollack (Eds.), *Intentions in communication*. Cambridge, MA: MIT Press.
Thomason, R.H. & Hobbs, J. R. (1997). Interrelating interpretation and generation in an abductive framework. In D.R. Traum (Ed.), *Working Papers of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, pp. 97–105. Menlo Park, CA: American Association for Artificial Intelligence.
Thomason, R.H., Stone, M. & DeVault, D. (2006). Enlightened update: a computational architecture for presupposition and other pragmatic phenomena. Ms., University of Michigan, Rutgers University, University of Edinburgh. Talk at the NSF-sponsored Workshop on Accommodation, OSU, October 13–15.
Tonhauser, J., Simons, M. Beaver, D. & Roberts, C. (2013). Towards a taxonomy of projective content. *Language*, 89(1), 66–109.

van der Sandt, R. (1992). Presupposition projection as anaphora resolution. *Journal of Semantics*, 9, 333–77.
van Kuppevelt, J. (1996). Inferring from topics: implicatures as topic-dependent inferences. *Linguistics and Philosophy*, 19(4), 393–443.
van Rooij, R. (2003). Conversational implicatures and communication theory. In J. van Kuppevelt & R. Smith (Eds.), *Current and New Directions in Discourse and Dialogue*, pp. 283–303. Dordrecht: Kluwer.
von Fintel, K.-U. (1994). Restrictions on quantifier domains. PhD dissertation, UMass, Amherst.
von Fintel, K. (2008). What is presupposition accommodation again? *Philosophical Perspectives: Philosophy of Language*, 22(1), 137–70.
Wilson, D. (1975). *Presupposition and non-truth-conditional semantics*. London: Academic Press.
Xue, J. & Onea, E. (2011). Correlation between presupposition projection and at-issueness: an empirical study. Proceedings of the ESSLLI 2011 Workshop on Projective Meaning, Ljubljana, Slovenia.

23

Lewis on Reference and Eligibility

J.R.G. WILLIAMS

This chapter outlines Lewis's favored foundational account of linguistic representation, and outlines and briefly evaluates variations and modifications. Section 23.1 gives an opinionated exegesis of Lewis's work on the foundations of reference – his *interpretationism*. I look at the way that the metaphysical distinction between *natural* and *non-natural* properties came to play a central role in his thinking about language. Lewis's own deployment of this notion has implausible commitments, so in section 23.2 I consider variations and alternatives. Section 23.3 briefly considers a buck-passing strategy involving fine-grained linguistic conventions.

23.1 Lewis's Interpretationism

23.1.1 Interpretationism

David Lewis was no fan of primitive intentionality. He wanted to explain how the intentional – including mental and linguistic representation – could arise in a fundamentally physical world. He would agree, I think, with Hartry Field:

> there are no "ultimately semantic" facts or properties, i.e., no semantic facts or properties over and above the facts and properties of physics, chemistry, biology, neurophysiology, and those parts of psychology, sociology, and anthropology that can be expressed independently of semantic concepts. (Field 1975, 386)

Lewis's *interpretationism* is one implementation of a more general strategy for reducing the representational properties of language: the *select-and-project* method. Selection: an entire semantic theory is singled out based on facts about linguistic practice in a given population. Projection: *what it is* for *N* to refer to *o* (for that population) is for the selected semantic theory to entail that *N* refers to *o*.

If the account is to do any work, then the selection step can't be described in terms of a correspondence of the claims of semantic theory with the semantic facts. Indeed, to get a reductive account, we'll need to cash out "selection" in a way that avoids mentioning semantic properties at

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

all. Interpretationists like Lewis give a two-step account of selection. The first component is to specify (in terms of the community's linguistic practices) a "target" pairing of sentences with semantic values. Lewis's own favored version of the first step pairs sentences with (coarse-grained) truth conditions; the pairing is established by appeal to conventional regularities in linguistic usage. Thus, if there is a convention among the relevant community of only uttering "la neige est blanche" when they believe that snow is white, the pairing between that sentence and the set of situations where snow is white is established. (Notice that Lewis appeals to an intentional relation – belief – in formulating the target pairing. This is in line with his "headfirst" methodology, whereby linguistic intentionality (reference, truth, etc) is reduced inter alia to mental intentionality (belief, desire, etc.); with a promise to give an independent account of the latter (cf. Lewis 1969, 1974, 1975, 1994).

Lewis's preferred view of the data should be distinguished from a different interpretationist account often associated with Lewis, *global descriptivism* (Lewis 1983, 1984). This requires pairing sentences with truth *values* rather than truth *conditions*. How to get a fix on the correlation? Perhaps every sentence that is treated by a community as "platitudinous" is paired with the true; and every sentence that is rejected offhand by the community is paired with the false. Or perhaps (Putnam-style) one pairs every sentence that is part of the idealized final total scientific theory with the true, and their negations with the false. The second component of interpretationism moves from data to selected theory. A natural constraint is: to be selected, a semantic theory must assign semantic values to sentences that fit the pairings on the target list.

To keep things simple, let's understand *fit* in the most naive fashion as the requirement that the semantic theory predicts a pairing of sentences and contents that exactly matches those that appear on the target list (for more sophisticated elaborations, see Lewis 1975). One minimal account of theory selection says that fit is necessary and sufficient for selection; more elaborate accounts impose further selectional constraints.

In the early 1980s, Lewis began advocating specific constraint on the selection of semantic theory over and above fit. He held that some theories were more "eligible" than others to be selected, because they assigned to lexical items more "natural" referents. This requires some explanation.

At the time he amended his interpretationism, Lewis had been convinced that for many tasks, throughout philosophy, an appeal to a distinction between *perfectly natural* and *merely abundant* properties was required (Lewis 1983). Paradigms of perfectly natural properties would be the fundamental notions of physics; paradigms of abundant properties would be ad hoc disjunctions such as Goodman's grue: *being green and first examined before 2100; or blue and not so examined*. Lewis took the distinction as more-or-less primitive. Perhaps a fan of Armstrongian universals could explain it in terms of those properties that are necessarily coextensive with a universal; perhaps we have a choice between positing primitive naturalness and positing a suitably rich contrastive primitive resemblance relation; but for present purposes this won't matter.

Lewis had some specific proposals about what properties had this status. Fundamental physics, he thought, would be our best guide to the perfectly natural properties instantiated at the actual world. Lewis claimed, furthermore, that this all-or-nothing division between the perfectly natural properties and the rest allowed us to make sense of a notion of relative naturalness – or more strongly, degrees of naturalness. The degree of naturalness of a property, says Lewis, is the minimal length of a definition of that property in perfectly natural terms. Thus, for example, if having positive charge and mass of 1 kg are each perfectly natural properties, having positive charge *and* unit mass might have a degree of naturalness of 1, where the degrees increase as naturalness decreases). For Lewis, semantic theories are graded as more or less eligible, depending on the naturalness of the semantic values that they assign to lexical items – the more natural (i.e. the shorter the definitional distance to the perfectly natural) the more eligible the theory. The selected semantic theory must optimize *fit* and *eligibility*.

Suppose that usage is equipoised between two semantic theories. For example, suppose that they assign exactly the same propositions to sentences, and so provide exactly the same content to feed into propositional speech acts such as assertion, questioning and commanding. Nevertheless, the symmetry may be broken, with one getting the semantic facts right. A foundational theory of language should pinpoint how the symmetry is broken. Some theories would appeal to internal relations between the vehicles and objects of content. Plato thought that some sounds were "names by nature," and were independently of usage apt to depict processes they resembled (a rolling r depicting rapid change, for example). Some theories would appeal to external relations, for example causal regularities between objects and the production of sounds. But an interesting alternative is to say that certain entities are "referents by nature" – the reason that they end up being named is that they are in themselves more nameable. Picturesquely, we call the phenomenon of symmetry-breaking with this source *reference magnetism*.

Lewis's appeal to eligibility predicts reference magnetism. Consider two semantic theories, differing only on the interpretation of one term *t* , which are equally good on grounds of fit. The first will be selected over the second when the candidate referent for *t* on the first is more natural than the candidate referent for *t* on the second, as this will make the overall theory more eligible. Eligibility is fixed by the array of referents, independently of any relation to words or the way they are used.

Why bring eligibility into interpretationism? In the next section, I'm going to argue that it arises naturally out of one tempting gloss on how selection should be conceived. But there are other benefits. A motivation for bringing in *something* other than fit into one's story about selection is that, in its absence the selection of semantic theory is simply too unconstrained. For example, if the only constraint on theory selection is that we generate the right pairing of sentences with truth values, then if a sensible seeming interpretation I generates the right pairing, we can construct a "permuted variant" I* which generates exactly the same distribution of truth values of sentences. Where the original theory said that "est blanche" picked out white things, the permuted variant might say that something was in the extension of that predicate iff its image under the permutation ϕ was white. So we overgenerate (radically! drastically! absurdly!) selected theories. Our account of selection-by-usage hasn't filtered down to anything like the intuitively credible candidates to be the semantic theory of a language. Still in the mix, for example, are interpretations on which my tokenings of "that puffin" refers to a small furry creature orbiting Alpha Centauri.[1] Lewis's eligibility constraint speaks to this concern. Even though the original and permuted variant are equally fitting – assign the same truth values or propositions – prima facie the permuted variant will be less eligible than the original – it takes slightly longer to spell out in perfectly natural terms. So eligibility (Lewis hopes) will allow us to resist the permutation inscrutability arguments and all sorts of other related difficulties and gives us the straightforwardly correct verdicts on what natural language terms refer to.

However, this is a weak motivation for introducing eligibility in particular, since all sorts of other symmetry-breaking considerations (for example, causal constraints) would do a similar job. Those who think that reference magnetism is a good description of semantic facts have more local and specific motivations. Theoretical terms are a nice source of test cases. Often, we want to view old-time theorists as saying insightful if sometimes false things about interesting subject matters, in preference to taking them to be talking about less interesting subject matters. Field (1973) makes a good case that the pre-twentieth-century account of mass was equipoised between rest mass, and the sum of mass and kinetic energy (what is sometimes called "relativistic mass"). My informants tell me that unqualified "mass" these days gets used for the former notion – it's regarded as more physically interesting, being independent of one's frame of reference, for example. Insofar as we want to say, in our present voice, that Newton talked and theorized about mass, we interpret him as talking about rest mass all along. While Field argued from equipoise in usage to referential indeterminacy in

old-timey deployments of "mass," we will regard this as a case in which somehow the symmetry is broken and Newton's term ends up determinately picking out *mass*. If the *reason* for this verdict appeals to the nature of the subject matter, then we're on board with reference magnetism, which Lewis's appeal to eligibility aims to capture.

23.1.2 Reference Magnetism and Naturalness

The presentation above is deliberately sloppy in a couple of respects. First, I gave Lewis's characterization of degrees of naturalness in terms of "definitional distance" of properties from the perfectly natural ones. But definitions are given in a certain language, and we don't pick out a language simply by listing a bunch of properties. Ted Sider (2011) suggests we respond to such concerns by extending the natural/non-natural distinction to entities of all categories, rather than just properties – in which case we can envisage a language "Ontologese" whose every bit of vocabulary stands for something perfectly natural. Whatever the merits of this, it goes beyond the resources available to Lewis himself. My interpretation of Lewis is as follows: we have a "canonical" language, which might include conjunction, negation, unrestricted first-order universal quantification, identity, perhaps plural quantification and mereological overlap, and of course the usual variable and punctuational symbols. Let's call these collectively the "auxiliary apparatus" of the canonical language. Aside from the auxiliary apparatus, the canonical language only contains predicates for perfectly natural properties. We can then say that P is definable in this language if there's an open sentence $\phi(v)$ of the language, such that necessarily, for all x, x has P iff x satisfies $\phi(v)$. However, semantic theories don't only assign semantic values to predicates; they also assign them to names, modifiers, operators, and so forth; and we may (indeed, I think we do) want to talk of the relative naturalness of the semantic values associated with these other kinds of lexical items. So we need an extended notion of definition to cover these cases. I'll assume we have one.

If we have the "canonical language" laid down, and some favored way of measuring the length of the definiens ϕ, then the notion of degree of naturalness can be taken to be the minimal length of such a definition – at least for those terms that have definitions at all. This is one locus for variation in a Lewis-style treat of eligibility, since it's not immediately obvious how lengths are to be measured, nor even what formal structure the "degrees" will take – will they induce a total or partial ordering? Ordinal or cardinal? Lexicographic or Archimedian? Even if we don't want to commit to a full theory of relative naturalness at this stage, fixing on formal features of the ordering is important. I will assume that the lengths are measured by integer values, so the ordering is total, cardinal and Archimedian (and with a natural zero). If you want a toy implementation, suggested by Lewis's writings, imagine that the length of ϕ is determined by counting the number of connectives that are present in ϕ. Notice that relations corresponding to the auxiliary apparatus – perhaps including identity and overlap – will be maximally natural, by this measure, even if they didn't appear on the list of perfectly natural properties and relations. Ultimately, these things will turn out to be "reference magnets" by the lights of Lewis's theory, just as much as the natural properties are. (If you feel queasy about this, and are prepared to engage in the additional meaty metaphysics, you might consider the Siderian alternative.)

But interpretationism does not appeal directly with relative naturalness. Instead eligibility – which I'm using as a term for some sort of ranking of semantic theories – is the primary concern. This is a second locus for variation within the Lewisian account – for even assuming that it's only the naturalness of properties that matters, we're being asked to move from a ranking of individual properties, to a ranking of semantic theories that assign many properties, of various degrees of naturalness. I propose we think of the degree of eligibility of a theory as the sum of the degrees of naturalness of the semantic values it assigns in the lexicon; but notice that this only makes sense because we

assumed that the initial comparative naturalness ranking assigned degrees that it makes sense to "add together." If we'd instead thought of the degrees as taking a partially ordered structure, this would not be available to us. We'll come back to this point below.

23.1.3 Best Grammar, Humean Simplicity, and Eligibility

We have the Lewisian version of select-and-project metasemantics on the table. It uses conventional regularities in linguistic usage to identify a target pairing of sentences with propositions, and then selects the semantic theory that optimally trades off fit against eligibility. The natural properties, because they make a semantic theory more eligible, are reference magnets. But presented in this way, the appeal to eligibility comes in somewhat from left field. It (perhaps) solves some problems, and its reference-magnetic predictions may be attractive. But is there anything more to say about why it shows up in an account of language? I think the deeper story, from which eligibility arises, turns on Lewis's handling of simplicity.

The appeal to natural properties only became a feature of Lewis's philosophy in the early 1980s. But his account of language pre-dates this. Here he is in 1975, responding to the puzzle we discussed above in connection to permutations – that fixing a function from sentences to propositions (which at this stage Lewis called a "language") underdetermines subsentential reference (part of what Lewis called a "grammar"). He has outlined a story about when a language counts as "in use" in a given population, and turns to the question about grammar-selection:

> Unfortunately, I know no way of making objective sense of the assertion that a grammar Γ is used by a population P, whereas another grammar Γ' which generates the same language as Γ, is not. . . .
>
> I do not propose to discard the notion of the meaning in P of a constituent or phrase, or the fine structure of the meaning of a sentence. To propose that would be absurd. But I hold that these notions depend on our ways of evaluating grammars, and are therefore no more objective than our notion of a *best* grammar for a given language.

Two key points to draw from this. (1) In 1975, Lewis's account of semantic theory ("grammar") selection, on the basis of sentence-level data ("language") is just that the selected theory is the best theory (where that is characterized by our standard evaluative methods). (2) He is worried that this story isn't "objective" – in particular, that by tying grammar-selection to evaluative methods the former will be "no more objective" than the latter.

Before moving on to tie this into eligibility, notice the immediate epistemic payoff of this early account of grammar-selection. The theorist of language arrives at judgments about what words mean by deploying certain evaluative methods. The truth about the subject matter is fixed (on this account) by the idealized application of those methods. So there's no mystery about how the epistemic methods of the theorist are appropriate to their subject matter. Theories of selection that say anything else will open up a skeptical gap – why are deployments of ordinary criteria good ways of finding out about semantic facts?[2]

One view of Lewis's development has him replacing the earlier account of selection with a novel condition – instead of speaking about the (worryingly non-objective) notion of best theory, we now appeal to the eligibility–fit trade-off. This would mean that his later self would have to address the epistemic challenge. But I think this is the wrong interpretation. There is continuity here, which we can see by looking at how Lewis tackles parallel issues in his Humean theory of laws.

Lewis Humean account of laws of nature said that the laws were the generalizations entailed by a "best" (optimally simple and informative) axiomatic theory. But the appeal to simplicity posed

problems. It's a familiar point that the (syntactic) simplicity of theories in general depends on the language in which they're formulated. To give Lewis's example: if we have a primitive predicate F, that expresses "being such that physical theory T holds," then the single axiom $\exists xFx$ will give us a maximally simple informational equivalent of T.[3] If by switching languages in this way we can achieve utter simplicity without sacrificing informativeness, then the simplicity component of the Humean theory loses traction, and the story collapses. In reaction, Lewis proposed that the relevant notion of simplicity operates on presentations of the theories concerned in a "canonical" language, built out of the perfectly natural predicates – indeed, exactly the canonical language relevant to assessing relative naturalness. It is syntactic complexity in this privileged representation that matters for Humean system selection – and the introduction of artificial predicates F is neither here nor there. Let's call the notion of simplicity so characterized "Humean."

Combine the 1983 account of (Humean) simplicity with the 1975 view of semantic theory selection. Take an axiomatic presentation of semantic theory *in the canonical language*. The complexity of the axiom assigning a semantic value to an expression *in this presentation*, will, modulo some constant factor, be given by the complexity of the definition of that semantic value in canonical terms (what we called earlier the degree of naturalness of that semantic value). Adding the complexities of axioms together gives the overall Humean complexity of the semantic theory we start with. However, the same calculation (modulo some constant) gives the overall eligibility of the theory. So Humean simplicity/complexity and eligibility/ineligiblity measure the same thing.

Lewis's interpretationism remains constant in form throughout. It is projectivist in a full sense of the word, with the typical epistemic payoffs – it's just that the account of simplicity is elaborated. The reference-magnetic behavior of more natural properties, and the resolution of radical inscrutability threats this provides, thus drops out of the underlying story.

23.1.4 Summary

Interpretationism has many virtues. Many attempts at a foundational account of linguistic representation are limited in scope. They might have stories to offer about the reference of names of our friends and the predicates for things we bump into and trip over – but this hardly scratches the surface of the vast motley of lexical items found in natural language. A total account of language needs to tell us about the meaning of modifiers, connectives, inflections, tenses, and much else besides. Interpretationism offers a single-shot solution to such troubles, since it gives a uniform story that grounds the fact that "very" means *very* just as well as it grounds the fact that "tiger" means *tiger* – in either case, because the facts are consequences of the selected semantic theory.

So long as we stick within the broad constraints of the 1975 framework on which selected semantic theory is best semantic theory, interpretationism offers a tractable epistemology. And so long as we can unpack the key notion of "best semantic theory" and characterize the data it is responsible to without circular appeal to linguistic representation, it offers a reduction of the semantic to the non-semantic. Quite a package!

One might worry that "best" semantic theory is just to nebulous or subjective to be appealed to in such a key role. Lewis's unpacking of this (as I see it) in terms of fit and eligibility extracts specific content from the gloss, in a way that allows us to pinpoint what goes wrong with indeterminacy/inscrutability arguments and underpins a reference magnetism that many find congenial as a description of how language works.

There is a huge amount to like about the view that has been built up. But I don't think it delivers on its promises; and while successor theories can follow through on some of the promises, I doubt any account will live up to the soaring ambitions of the original.

23.2 Credible Reference Magnetism

The above discussion talked as if we get sensible results out of a Lewisian metasemantics, on which eligibility is analyzed ultimately in terms of definitional distance from the perfectly natural – which in Lewis's case, meant microphysical properties. This is a problem.

The problem is rather straightforward. I don't really have a clue what a "definition" of the ordinary subject matter of thought and talk – shoes and string and sealing wax – would be, if the definiens is to be drawn from microphysics. I suspect you don't either. There's no reason to believe that such definitions are even possible. Incredulous stares at this presupposition of the Lewisian package are common. I think that the incredulity can be turned into specific arguments against the account (see Williams 2007a).

I won't argue any further against the Lewisian proposal here – so I'll leave it open for others to make the case that ordinary notions are finitely definable in the required way, and that the arguments that it leads to trouble can be blocked. Instead, I will work with the assumption that the Lewisian proposal as originally envisaged fails, and examine what prospects remain for an eligiblity-based interpretationism.

I'll look at three proposals: two that revise the overall metaphysics, and one that keeps the metaphysical framework intact but drops the connection between eligibility and naturalness.

23.2.1 Response 1: Macronaturalism

The troubles for Lewisian eligibility, it might be thought, do not originate from the theory of eligibility *per se*, nor in its relation to perfect naturalness. The worries stems from the fact that Lewis commits to the view that (in the actual world) the perfectly natural properties are to be found in microphysics, and not in the "macroworld." But, one might argue, wherever we find law-like connections; wherever we find genuine objective similarity; and wherever we find causation, we should believe that we're working with perfectly natural properties. And, it may be argued, we encounter such phenomena in the macroworld of geology, biology, and ecology as much as the microworld of theoretical physics (Schaffer 2004). The perfectly natural properties will be sparse but not ultra-sparse.

Whether eligibility-based interpretationism remains reductive in this setting is open to question – if we're allowing in the equivalent of biological and ecological universals, what about psychological ones – beliefs and desires? Indeed, what about the special science of semantics? One might think that a principled version of this macroworld picture should include perfectly natural relations corresponding to intentional verbs, or even reference itself.

Even if we can take it that the vocabulary of the special sciences is available to us, it's not clear how we'd go about defining terms for artifactual kinds, nor the variety of verbs we use in everyday life (terms for thick ethical or aesthetic concepts, for example). So even if the range of resources we have available as a definitional base isn't as recherché as on the Lewisian proposal, the definitional ambition is still grandly ambitious. Of course, the reductive achievement in prospect, and the definitional ambition, play off one another: the more sparse a macroworld we buy into, the more reductive the final proposal, but the bigger the definitional task we set ourselves.

The macroworld view of the distribution of perfectly natural properties raises questions about what the wider role of natural properties is to be. One idea that is prominent among Lewisians is that natural properties should enjoy a certain kind of modal independence. Chalmers's dualism (Chalmers 1996) is a natural illustration of the sort of thing we might expect: on this kind of view, if some physical thing is conscious, it's possible for there to be a physical duplicate of it that is not a duplicate *simpliciter*, because it lacks consciousness. One question for the macronaturalist is whether

something similar goes for their properties. Is it possible to have physical duplicates that are not chemical or biological duplicates, for example? If not, aren't we committed to some objectionable "necessary connections between distinct properties"? Again, if special science kinds (for example) are perfectly natural, then won't the vagueness and indeterminacy of (e.g.) biological kinds give us "vagueness in the world"?[4] The macronaturalistic view demands we revise our entire conception of the fundamental structure of the world to support a metaphysics of one very special part of the world – linguistic representation. Is a more localized and modular response to our problems available?

23.2.2 Response 2: Comparative Naturalism

A second response to worries about definability from microphysics is that Lewis went wrong (or that I went wrong in interpreting Lewis) by attempting to reduce relative naturalness to perfect naturalness via "lengths of definitions" from some canonical language. Maybe we should stick with relative naturalness itself as primitive. After all, the motivating cases for this distinction are examples that involve comparative judgments very distant from what Lewis regards as perfectly natural: that green is more natural than grue; that *being the image under a permutation of something human* is less natural than being human; that artifactual kinds are less natural than biological kinds – and so forth. More generally, the idea of a property P being "grounded" or "holding in virtue" of Q is fairly widespread in contemporary metaphysics, for better or worse – and one might think that this notion gives rise directly to comparative naturalness at all levels of reality. If comparative naturalness is rock bottom, then an addition bonus (some may think) is that we can be agnostic over whether there is a layer of "maximally natural" properties in the first place opens up – perhaps there are simply more and more natural properties ad infinitum (compare Langton and Lewis 1998; Schaffer 2003).

Just as with macronaturalism, we need to consider how primitive comparative naturalism integrates into wider theory. For example, it might sound attractive to countenance the possibility of ever-more-natural properties; but if the theoretical deployments of naturalness appeal to the all-or-nothing concept, then it's not clear that comparative naturalness will be an adequate replacement. As an illustration: Lewis's original theory of duplication and intrinsicality made appeal to the sharing of perfectly natural properties (cf. Lewis 1983, 1986). The theory of Langton and Lewis (1998), which proved far more problematic, is one exactly designed to liberate the analysis of intrinsicality from appeal to a layer of maximally/perfectly natural properties. That it runs into worries that do not face the original account illustrates the damage such shifts in resources can inflict.

There are more local concerns about the adequacy of an appeal to primitive comparative naturalness in connection to metasemantics. Recall that earlier we emphasized the distinction between relative naturalness (of properties) and relative eligibility (of whole theories). Let's think of a toy case: an object language that has the syntax of first-order logic, with only three non-logical terms, the predicates "rock," "tree," and "human." To cut down on complexity, look at semantic theories that differ only over their assignments to the non-logical predicates. Compare three candidate assignments to "rock," "tree," and "human," respectively: $A = (A_1, A_2, A_3)$, $B = (B_1, B_2, B_3)$ and $C = (C_1, C_2, C_3)$. We may suppose, writing $>$ for "more natural than," we have: $A_1 > B_1 > C_1$, $B_2 > C_2 > A_2$ and $C_3 > A_3 > B_3$. This component-wise ranking gives us three different induced rankings of the overall assigned, which display a cyclic Condorcet-structure: A beats B twice, B beats C twice, and C beats A twice. In that kind of situation, how are we to rank the candidate assignments for relative eligibility?[5]

We earlier pointed out that one question for the theorist of comparative naturalness faced was over the formal characteristics of the ordering. Partial or total? Cardinal or ordinal? Lexicographic

or Archimedian? A zero element or not? On some ways of answering this question, there will be natural ways of sorting out Condorcet-style troubles – for example, in the Lewisian account that (I took it) delivered integral degrees of naturalness, we could simply look at the sum total of the various degrees. But that kind of structure seemed reasonable because we were analyzing comparative naturalness in terms of a metrical notion – lengths of definitions. In the current context, there's no reason yet to think that summing "degrees" of naturalness makes sense (particularly if they're partially ordered, or have no natural zero). And of course, the richer the structure we're forced to posit, the more costly it is to take it as explanatorily basic. The moral from these discussions is that we need to keep sharply in view that the resource that interpretationism needs is a ranking of whole theories, not individual properties. Without a story of how to get from the latter to the former, we have no theory at all.

23.2.3 Response 3: Parochial Eligibility

The final option I will consider involves dropping the link between naturalness and eligibility, but retaining the rest of the structure of eligibility-based interpretationism. The story developed earlier can be seen as giving a programmatic specification of degrees of eligibility. We input some "canonical language" L. We obtain an ordering of properties (and other entities) by minimum length of definitions in L – and by summing we thereby obtain what we might call "degrees of eligibility$_L$" attaching to sets of entities/whole theories. Degrees of eligibility$_L$ are then traded off against fit to select the meaning-fixing theory of interpretationism. So each choice of L delivers an account of reference – reference$_L$. Question: for what L can we plausibly maintain that reference *simpliciter* is reference$_L$?

There are many candidates for L. Lewis (and others following him) propose some kind of construct drawn from fundamental metaphysics ("Ontologese"), which we've seen causes problems. But L could be, for example, English. After all, the metalanguage of semantic textbooks tends to be a natural language (suitably supplemented by technical vocabulary) – not some artificially restricted language drawn from metaphysics or other special sciences. If we start with English, we have something with the recognizable pattern of eligibility-based interpretationism, but the base from which eligibility is determined is parochial rather than metaphysical.[6]

But if we choose the wrong L, the whole project could be threatened. Consider what one might say against the proposal of taking L = English (an extreme example, but a useful stalking horse). An overarching concern is that we are in effect offering a listiform, disjunctive characterization of what it is to be an eligibility-maker: *x* is an eligibility-maker iff *x* is either *being a shoe* or *being string* or *being sealing-wax* or . . . and so on, adding a disjunct for each entry in the English dictionary. But this seems an utterly ad hoc and arbitrary foundation to build a theory of reference on – why English rather than French or Mandarin? Why English-as-it-actually-is rather than English-as-it-might-well-have-been?

There is a host of more specific worries.

- First, the ambitions of the appeal to eligibility might be unrealized. It was supposed to be reductive: but now reference is reduced in part to eligibility, and eligibility is defined, inter alia, via appeal to reference, truth and other semantic relations that English has words for. Moreover, part of the appeal of eligibility is that it was predictive: given the verdict that rest mass is more natural than rest-mass-plus kinetic energy, then we generate a prediction that an inchoate use of "mass" picks out rest mass. But if English has a single lexical item "mass," then all we get is the uninformative statement that the reference magnet in the vicinity of "mass" is mass (whatever that is!)

- Second, the proposal looks extensionally inadequate. Suppose the French apply "vert" roughly in the way English speakers apply "green," but slightly shifted to the bluish end of the spectrum. Then it's very plausible that the extension of this color term is slightly blue-shifted, compared to the English color term "green." But if the properties for which we have English terms are reference magnets, then our account would presumably predict that "vert" picks out greenness – a false prediction. A related problem is that some words in English are so flexible that the whole account threatens to trivialize, leading us back to permutation problems or worse. If we can use the demonstrative "that" to pick out some permuted variant of green, then a permuted interpretation of natural language could be phrased by simple demonstratives in the semantic axioms to assign a maximally "eligible" permuted semantic theory.
- Third, even if the account is extensionally adequate (getting the right actual meanings to actual words) it goes wrong in counterfactual scenarios. Consider the Ectoplasmians, living in an environment none of whose most important characteristics we lack words for. It seems possible for alien creatures in alien environments to refer to things around them. But there's no reason to believe an interpretationism parochially tied to the things *we* have words for as reference magnets, will give decent results in such far-off scenarios.

This is a long list of objections and worries. Yet I think the parochial strategy is actually the most promising of those considered so far. The three kinds of specific worries just raised are very sensitive to the particular choice of *L* we considered – and this was just a stalking horse. If one refined, supplemented and improved the English language before using it as the basis for an analysis of Eligibility, many of the problems of detail could be finessed. Likewise if we input not the whole of English but some interesting principled core (scientific English, or the final vocabulary of idealized science, perhaps). And it's not so clear how serious the complaints are, on reflection – if other concerns were addressed, I would accept a putative reduction of reference that gave counterintuitive results only for very alien environments.

But even if the worries of detail can be finessed, the overarching worry of principle remains: at the heart of parochial eligibility is the reduction of reference to a notion cashed out by a long, arbitrary, ad hocly chosen disjunction. Field (1972) famously objected to putative reductions of reference on grounds of hidden disjunctiveness; one might think that the project immediately loses interest if it has to appeal to such things. One might indeed think of the parochial strategy as a kind of semantic deflationism – but one that does not deflate semantic relations directly (e.g. via the disquotation schemata) but instead goes deflationary on a notion in the reductive base – Eligibility.

Does the underlying disjunctiveness mean the view is a non-starter? I think we here need to draw on some of the debates that contemporary deflationists have prompted.[7] At least some relations and properties are such that disjunctive reductions are obviously correct (consider grue). So what we need to think about is the conditions under which disjunctive reductions are legitimate, and figure out whether reference falls inside or outside the category. A plausible suggestion is that disjunctive reductions are unavailable if the property to be reduced has an autonomous explanatory role. The tenability of the parochial *strategy* would then be tied to questions about whether semantic relations have an explanatory role, and if so, which semantic relations do so. Even if truth conditions have an important theoretical role to play, it's not clear that subsentential reference would thereby feature ineliminably in explanations.

If nothing else, studying parochial treatments of Eligibility serves to highlight some of the benefits of Lewis's original suggestion of Ontologese as the canonical *L* in terms of which interpretationism is to be understood – and thus allows us to identify features that a plausible successor should preserve.[8]

23.3 Buck-passing

Hitherto, I've been assuming with Lewis that the data to which the best grammar is responsible is a pairing of sentences with coarse-grained propositions. Even to get so far, Lewis has to endorse a "headfirst" methodology whereby linguistic content is analyzed, not directly in terms of the non-intentional, but in terms that presuppose mental intentionality – the (coarse-grained) beliefs and desires that enter into his analysis of convention. So there's an element of buck-passing built into Lewis's account: propositional content is inherited from mental content via conventions; it is objectual content (and similar) that is the distinctive novelty at the linguistic level. The burden of introducing objectual content into the picture falls on the account of semantic theory selection.

But there's nothing in the overall shape of Lewisian interpretationism that rules out a more thorough buck-passing. For example, unlike Lewis, we might help ourselves to fine-grained belief contents. We could have very fine-grained conventions – a convention regularity of assenting to "that square is red" only when one has a belief whose content is the fine-grained proposition with first component a certain square, and second component redness. Such a rich set of sentence-proposition pairs would not allow much wriggle room for permutation style arguments.[9]

Can such fine-grained conventions associating each sentence with what is in fact the structured proposition they express? And can one articulate them in a way consist with an overall reductive project? To answer this, one must take a stand on several points of detail:

1. What kind of story is in prospect for reducing fine-grained mental content? Is the story interpretationist, or of some other kind? (Presumably it won't be convention-based interpretationism; but some kind of global descriptivism run on mentalese may yet be an option).
2. How rich is fine-grained mental content? Does it allow properties remote from direct experience to feature as constituents (electronhood)? Does it contain constituents corresponding to the features of natural languages: tense, aspect, case, modifiers, intensifiers, etc?
3. Is all fine-grained mental content reducible independently of linguistic content (and hence available for a non-circular reduction of the latter), or just a fragment of it?

Suppose that: (i) a *restricted* range of fine-grained mental content was available to feed into fine-grained conventions associating, for example, "the patch is white" with the structured proposition *that the patch is white*; but (ii) only for sentences with macroscopic or phenomenal subject matters can we *non-circularly* spell out the story of the determination of the fine-grained mental content. If so, a fine-grained convention associating "the electromagnetic way is high frequency" with belief in the structured proposition thereby expressed – even if it exists – couldn't form part of a *reductive* account of linguistic content. Where would this leave the buck-passing project?

One should bear in mind that the interpretationist story about "fit" *must* be fairly sophisticated, to cope with everday phenomenona such as metaphor, implicature, loose speech, exaggeration, and so forth. Fitting data isn't just a matter of the semantic theory pairing a sentence with the proposition conventionally associated with it – instead, the semantic theory must explain that conventional association, modulo general principles of pragmatics, facts about epistemic limitations, and so forth. Second, even if there are no fine-grained conventions associating "the electromagnetic ray is high frequency" with the structured proposition it expresses, there may yet be a fine-grained description of default macroscopic/phenomenal verification conditions. A semantic theory that associates electromagnetic rays with "electromagnetic ray" may still count as fitting that fine-grained pairing since, on the hypothesis that *S* means *p*, you'd expect to find a regularity of assenting to *S* when one believes *q*, when *q* articulates canonical evidence for *p*.[10]

The buck-passing strategy is no panacea. First, we have seen there are several ways of generalizing the Lewisian appeal to conventions to exploit fine-grained content, and it's unclear which is best. Second, it doesn't remove the need for hard thinking about constraints other than *fit* in selection of best grammar. For example, whether conventions are fine-grained or coarse-grained, as regularities in actual usage there are presumably only finitely many of them, and this leaves room for the variations on Kripkensteinian challenges that Hawthorne (1990) presses. For each "sensible" semantic theory there are "twisted" variants differing from it on the truth-conditions assigned to (grammatical) sentences which are never used. Lewis (1992) invokes naturalness/eligibility in response. Finally, the buck-passing strategy by its nature shifts the burden of providing an account of the determination of objectual intentionality from the linguistic to the mental, and so we are always involved in a trade-off between the resources we help ourselves to and the hostages to fortune we give.

Nevertheless, buck-passing this might be a productive relocation of the problem of objectual intentionality. Different forms of representation can have very different characteristics. While the relationship between the term "square" and squareness is paradigmatically extrinsic, the relationship between percepts and their contents seems rather different . It seems no coincidence that squarish percepts represent squares – at first pass, this seems like an internal relation between the vehicle of representation and its content, in sharp contrast to the linguistic case (hence the tradition, represented by Locke, on which the internal relation of resemblance is assigned a primary role in the analysis of mental intentionality – a far more plausible suggestion than Plato's appeal to a similar word–world resemblance constraint). In passing the buck, we don't just shift the problem of determining fine-grained content around – we relocate it to a position where both constraints and resources for tackling it are quite different.

23.4 Conclusion

While the account that Lewis gives has a lot to recommend it in the abstract, it is problematic.

One might stick with the letter of Lewis's 1980s revision/supplementation to his interpretationism – but then no one should have very high initial confidence that the theory is even extensionally adequate. One could go for one of the variants above – the superheavy metaphysics of macronaturalism, the seriously underdeveloped theory of comparative eligibility, or the defanged account of eligibility divorced from naturalness that characterizes the parochial strategy. There are plenty more variations and alternatives! But appeal to eligibility without some indication about the intended framework is not a wise option for the conscientious reference-reducer.

What might the best successor to Lewisian interpretationism look like? I've emphasized throughout the sheer interconnectedness of the theory of linguistic content – it interacts in complex ways with metaphysics, the theory of mind, the explanatory role of content, and so on. In such cases, verdicts reached before the philosophical endgame must be tentative. But as we finish, it is worth picking out one candidate route through the maze of options.

Much of the Lewisian framework remains attractive. The overall select-and-project method, with its promise to deal even-handedly with all linguistic intentionality, and its epistemic tractability, should be retained. As I see it, this means that we should work within the constraints of the 1975 picture whereby the selected semantic theory is the "best theory" of a certain range of data. And, as Loewer (2007) urges in the analogous case of Humean laws, I think that what we say about "best theory" should be informed by looking at what aspects of theories are treated as theoretical virtues by practitioners. I think it plausible we will retain the eligibility-type structure, but I think it's very implausible that metaphysics will provide a canonical language to feed into this structure.

A modicum of buck-passing will be possible, I think: certain kinds of fine-grained contents will be available to be appealed to, so that *pace* Lewis, fine-grained conventions could be data for interpretationist theory selection. For a core fragment of language, this fine-grained data may pair sentences with the very thoughts or structured propositions they express. For more recherche′ vocabulary, I doubt this will be the case. We will either have to retreat to coarse-grained conventions, or fine-grained conventions that connect sentences to something like verification rather than truth conditions. As emphasized above, this data could still select for a standard semantic theory – but (together with Kripkensteinian worries mentioned earlier) it does mean that buck-passing can't resolve all our problems.

So we return to the account of eligibility. I see the challenge much as it is set up in the section on parochial eligibility above: for which choice of canonical metalanguage *L* will the theory return plausible results? The big choice point is whether to go for something at least semi-principled (and so face the hard task of articulating it and explaining what is principled about the choice made) or go for something utterly parochial (and so face the hard task of arguing that the lack of principle doesn't matter).

The choice here should be guided by one's views on the explanatory role of content. Contemporary deflationists sometimes argue that linguistic content (whether lexical or sentential) has no autonomous explanatory role. This would remove the main objections I see to the parochial strategy above – and I think a parochial form of interpretationism has much to recommend it over a straightforward deflationism about truth and reference (for example, it says much more plausible things about local modal profile of the truth values of sentences).

I don't really believe the deflationist claim about the explanatory redundancy of linguistic content. But I am impressed by the comparative prominence of appeals to *sentential* content in linguistic explanation (e.g. in analyzing the effects of speech acts like assertion, questioning, commanding, etc.). Conjecture: the only explanatory role for linguistic content involves the truth conditions of sentences (so that reference, if mentioned at all, only features to explain why sentences have the truth conditions they do).[11] This is a substantive, highly non-trivial claim. If it were correct, I'd be inclined to defend a sophisticated parochial treatment of eligibility ("sophisticated" because it must deal with the problems of detail identified earlier). The earlier appeal to fine-grained conventions constraining fit will help to make the case that a wider class of linguistic facts will be invariant under various parochial choices of *L* than one might fear.

This is one way among many that a post-Lewisian interpretationism might head. It results from emphasizing and retaining some of the virtues of Lewis's original proposal promises (e.g. universality and epistemic tractability), while compromising on others (e.g. metaphysical principle). No doubt other combinations of retainment and compromise are possible. One very interesting question is whether there are any credible developments of the view that deliver on all the promises of Lewis's original. But even if we don't get this, the compromising descendants stand on their own as theories of the foundation of linguistic content.

Notes

1 For some of the original literature on this 'inscrutability of reference' problem see Jeffrey 1964; Quine 1964; Wallace 1977; Davidson 1979; Putnam 1981. See Williams (2008c) for a general version of the permutation argument in application to rich languages, and a discussion of why inscrutability is to be avoided for theoretical as well as intuitive reasons.

2 The epistemology of the theorist, at least, is well accounted for. The first-person epistemology of the in situ language *user* is not so clearly addressed – but I won't consider this tricky issue here.

3 At least, it does if we assume that informational equivalence is defined in coarse-grained terms. I don't want to pursue this further here.
4 For some exploration of life with a primitive macroworld, see Williams 2007b, 2008d. For vagueness in the world, see Williams 2008a; Barnes and Williams 2011.
5 Following up the analogy to voting paradoxes, it's worth asking whether a version of Arrow's theorem lurks in the vicinity. Pareto and non-dictatorship constraints are certainly plausible. However, it's not clear to me whether there's an interpretation of the Universality and Independence axioms that allows us to invoke the theorem in this setting. (Addendum: Hansen and Morreau, following up on the suggestion in this note, argue that a version of Arrow does indeed pose problems here.)
6 Compare Davidson (1973) – Davidson's radical interpreters speak some particular natural language, and this kind of identification of the eligibility-basis introduces features reminiscent of Davidsonian radical intepretation into an otherwise Lewisian version.
7 Field (2001) contains a number of papers linking the debate between inflationism and deflationism to the explanatory role of content.
8 The parochial strategy is a vegetarian cousin of Taylor's notion of '*T* -cosy' properties (Taylor 1993). But whereas Taylor proposed relativizing key notions to theories (e.g. eligibility, and consequently reference) I am considering the prospects of derelativizing in some parochial manner. An interesting comparison is Loewer (2007) on Humean laws, who suggests selecting for language and theory together. If we could select 'best language' as well as 'best theory' in some principled way, this would be a most attractive option.
9 The reasons are much the same as in Williams 2008b.
10 One is walking a fine line here, of course. A permuted semantic theory may make explicable fine-grained conventional regularities, given minimal competence in unwinding the permutations.
11 Compare Field 1975.

References

Barnes, E.J. & Williams, J.R.G. (2011). A theory of metaphysical indeterminacy. *Oxford Studies in Metaphysics*, 6, 103–48.

Chalmers, D. (1996). *The conscious mind*. Oxford: Oxford University Press.

Davidson, D. (1973). On the very idea of a conceptual scheme. *Proceedings and Addresses of the American Philosophical Association*, 47, 5–20. Reprinted in D. Davidson (1980), *Inquiries into truth and interpretation*, pp.183–99. Oxford: Oxford University Press.

Davidson, D. (1979). The inscrutability of reference. *Southwestern Journal of Philosophy*, 10(2), 7–19. Reprinted in D. Davidson (1980), *Inquiries into truth and interpretation*, pp. 227–42. Oxford: Oxford University Press.

Field, H.H. (1972). Tarski's theory of truth. *Journal of Philosophy*, 69, 347–75. Reprinted in Field (2001), pp. 3–29.

Field, H.H. (1973). Theory change and the indeterminacy of reference. *Journal of Philosophy*, 70, 462–81. Reprinted in Field (2001), pp. 177–98.

Field, H.H. (1975). Conventionalism and instrumentalism in semantics. *Noûs*, 9, 375–405.

Field, H.H. (2001). *Truth and the absence of fact*. Oxford: Oxford University Press.

Hawthorne, J. (1990). A note on languages and language. *Australasian Journal of Philosophy*, 68(1), 116–18.

Jeffrey, R. (1964). Review of *Logic, methodology and the philosophy of science*, ed. E. Nagel, P. Suppes & A. Tarski. *Journal of Philosophy*, 61, 79–88.

Langton, R. & Lewis, D. (1998). Defining "intrinsic." *Philosophy and Phenomenological Research*, 58(2), 333–45.

Lewis, D.K. (1969). *Convention: a philosophical study*. Cambridge, MA: Harvard University Press.

Lewis, D.K. (1974). Radical interpretation. *Synthese*, 23, 331–44. Reprinted in D. Lewis (1983), *Philosophical papers*, Vol. I, pp. 108–18. Oxford: Oxford University Press.

Lewis, D.K. (1975). Language and languages. *Minnesota Studies in the Philosophy of Science*, 7, 3–35. Reprinted in D. Lewis (1983), *Philosophical papers*, Vol. I, pp. 163–88. Oxford: Oxford University Press.

Lewis, D.K. (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, 61, 343–77. Reprinted in D. Lewis (1999), *Papers on metaphysics and epistemology*, pp. 8–55. Cambridge: Cambridge University Press.

Lewis, D.K. (1984). Putnam's paradox. *Australasian Journal of Philosophy*, 62(3), 221–36. Reprinted in D. Lewis (1999), *Papers on metaphysics and epistemology*, pp. 56–77. Cambridge: Cambridge University Press.

Lewis, D.K. (1986). *On the plurality of worlds*. Oxford: Blackwell.

Lewis, D.K. (1992). Meaning without use: Reply to Hawthorne. *Australasian Journal of Philosophy*, 70, 106–10. Reprinted in D. Lewis (1999), *Papers on ethics and social philosophy*, pp. 145–51. Cambridge: Cambridge University Press.

Lewis, D.K. (1994). Reduction of mind. In S. Guttenplan (Ed.), *A companion to the philosophy of mind*, pp. 412–31. Oxford: Blackwell. Reprinted in D. Lewis (1999), *Papers on metaphysics and epistemology*, pp. 291–324. Cambridge: Cambridge University Press.

Loewer, B. (2007). Laws and natural properties. *Philosophical Topics*, 35(1&2), 313–28.

Putnam, H. (1981). *Reason, truth and history*. Cambridge: Cambridge University Press.

Quine, W.V. (1964). Ontological reduction and the world of numbers. *Journal of Philosophy*, 61. Reprinted with substantial changes in W.V. Quine (1976), *The ways of paradox and other essays: revised and enlarged edition*, pp. 212–20. Cambridge, MA and London: Harvard University Press.

Schaffer, J. (2003). Is there a fundamental level? *Noûs*, 37, 498–517.

Schaffer, J. (2004). Two conceptions of sparse properties. *Pacific Philosophical Quarterly*, 85, 92–102.

Sider, T. (2011). *Writing the book of the world*. Oxford: Oxford University Press.

Taylor, B. (1993). On natural properties in metaphysics. *Mind*, 102(405), 81–100.

Wallace, J. (1977). Only in the context of a sentence do words have any meaning. In P.A. French & T.E. Uehling Jr. (Eds.), *Midwest studies in philosophy 2: studies in the philosophy of language*. Morris: University of Minnesota Press.

Williams, J.R.G. (2007a). Eligibility and inscrutability. *Philosophical Review*, 116(3), 361–99.

Williams, J.R.G. (2007b). The possibility of onion worlds. *Australasian Journal of Philosophy*, 85(2), 193–203.

Williams, J.R.G. (2008a). Multiple actualities and ontically vague identity. *Philosophical Quarterly*, 58(230), 134–54.

Williams, J.R.G. (2008b). Permutations and Foster problems: two puzzles or one? *Ratio*, 21(1), 91–105.

Williams, J.R.G. (2008c). The price of inscrutability. *Noûs*, 42(4), 600–41.

Williams, J.R.G. (2008d). Working parts. In R. Le Poidevin (Ed.), *Being: contemporary developments in metaphysics*. Royal Institute of Philosophy Supplement, Vol. 83. Cambridge: Cambridge University Press.

24

On the Nature of Certain Philosophical Entities

Set Theoretic Constructionalism in the Metaphysics of David Lewis

GIDEON ROSEN

24.1 Lewis's Ontological Scheme

Viewed from a suitable distance, David Lewis's ontological scheme is simplicity itself. Absolutely everything that exists, according to Lewis, is either a spatiotemporal particular, or a set theoretic construction from such particulars, or a mereological aggregate of such items.[1] Entities that other writers treat as *sui generis* – properties, relations, events, propositions, possible worlds and individuals, mental contents, languages, linguistic meanings – are all either identified with concrete things (in the case of possible worlds and individuals) or with set theoretic constructions therefrom. Thus Lewis identifies properties with arbitrary classes; relations with classes of ordered pairs, propositions with classes of possible worlds (or structured constructions from possible individuals), events with certain classes of spacetime regions, and so on.

Lewis inherited this scheme from Quine.[2] Lewis and Quine disagree about the inventory of particulars, and perhaps about the principles of set theoretic construction. But they agree about this much: when God made the world(s), he made the concrete individuals; he laid down the principles of set theory, and then he stopped.

Set theoretic constructionalism is not an incidental feature of Lewis's system. The master argument of *On the Plurality of Worlds* is that a pluriverse composed of infinitely many concrete universes constitutes a "paradise for philosophers" wherein "we find the wherewithal to reduce the diversity of notions we must accept as primitive, and thereby to improve the unity and economy of the theory that is our professional concern – total theory, the whole of what we take to be true" (Lewis 1986a, 4). But as the argument develops it becomes clear that this is not an argument for modal realism on its own. It is an argument for a package that includes modal realism *and set theory*, together with a system of reductive definitions that identify various "philosophical entities" (as we may call them, following Montague) with set theoretic constructions from possibilia. Without set theory, the pluriverse is not a paradise, and, if it is not a paradise, we have no reason to accept it.

After the completion of *Plurality*, which is mainly concerned with the possibilia that form the basis for these constructions, Lewis turned his attention to the metaphysics of sets and classes. Given the

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

role of set theory in mathematics, this is a compulsory topic on any view. Any comprehensive metaphysical theory must include a chapter about the nature of sets and the relation of set membership. But for writers like Lewis and Quine (and Carnap and Russell) who see set theoretic construction at work not just in mathematics but throughout philosophy, the topic is especially important. It is therefore striking that in his writings on these topics, Lewis invariably treats the philosophy of set theory as a special topic in the philosophy of mathematics. Given his other views, his account of classes has profound implications for general metaphysics, implying as it does an account of the nature of properties and instantiation, propositions and propositional truth, the nature of events, and so on. But Lewis does not consider those implications.[3] The aim of this chapter is to close the circle by bringing Lewis's mature account of sets into contact with his larger ontological scheme. As we shall see, the results are sometimes surprising.

24.2 Metaphysics of Classes: The Simple View

Lewis's mature views on sets are found in the short monograph *Parts of Classes* (1991) and the follow-up paper "Mathematics is Megethology" (1993). In fact, however, these works present two rather different positions. The body of *Parts of Classes* develops what I shall call the *Simple View of classes*, whereas the appendix, co-authored with John Burgess and Alan Hazen, and "Mathematics is Megethology" develop an alternative, *Structuralist View*. In this section and the next, we briefly review these two positions.

Lewis begins from the thought that like ordinary particulars, classes have mereological (part–whole) structure. Just as your car is the sum of its parts – carburetor, windshield, etc. – so the class $\{a, b\}$ is the sum of *its* parts. What are the parts of this set? It is tempting to suppose that $\{a, b\}$ is somehow made of a and b, and more generally, that the parts of a class are its members. But on reflection, this can't be right.[4] Let a and b be mereological atoms – things without proper parts – and let c be their *fusion*: the item that contains a and b as parts and nothing else. Now consider the classes $\{a, b\}$ and $\{c\}$. These are clearly different: the first has two members, the second one. But if the parts of a set are its members, then – given standard assumptions about parts and wholes – they must be identical. For given this hypothesis, $\{a, b\} = a + b$, and $\{c\} = c = a + b$, from which it obviously follows that $\{c\} = \{a, b\}$.[5]

So we cannot say that the parts of a class are its members. Instead we should say:

(Main Thesis) The parts of a class are its (non-empty) *subclasses*.[6]

On this view, $\{a, b\}$ has two proper parts. But they are not the individuals, a and b. Rather they are the *singletons* $\{a\}$ and $\{b\}$. The set $\{c\}$, by contrast, has no proper parts, since it has no non-empty subclasses besides itself.

The Main Thesis entails that singletons are mereological atoms. Pick some item, say, your car. Set theory guarantees that it has a singleton, {car}. But despite the notation, there is no sense in which this singleton is *constructed from*, or *made up out of* your car. The singleton is not made out of anything. For all we have said, {car} might be part of your car, or part of the Crab Nebula. But your car is certainly not part of *it*.

In the body of *Parts of Classes*, Lewis takes the notion of singleton as the sole primitive notion of set theory. He then defines the main concepts of set theory as follows:

An *individual* is a thing with no singletons as parts.
The *null set* is the fusion of the individuals.[7]

A *class* is a fusion of singletons
X is a *member* of *Y* iff *Y* is a class and the singleton of *X* is part of *Y*.

The central mathematical result of the book is that given these definitions, all of standard set theory (and a bit more) can be recovered by combining a bold hypothesis about the number of atoms[8] with classical mereology and a theory of the singleton relation. The details will not matter, but it will be helpful to have this theory in front of us:

Functionality:	Everything has at most one singleton
Distinctness:	No two things have the same singleton
Domain:	Every individual has a singleton
	Every singleton has a singleton
	Every small fusion of singletons has a singleton[9]
	Nothing else has a singleton
Induction:	If there are some things such that
	• every individual is one of them
	• every singleton is one of them
	• every fusion of some of them is one of them
	then everything is one of them.

When this theory of singletons is combined with classical mereology and the hypothesis that the universe contains inaccessibly many atoms, the result is second-order Morse-Kelly set theory with individuals, an extension of Zermelo-Frankel set theory that is sufficient for almost all known mathematics, pure and applied.[10]

This reaxiomatization permits us to isolate the philosophically problematic parts of set theory. Classical mereology is perfectly unproblematic, as Lewis argues. But, on Lewis's account, all of the *collecting* and *combining* and *constructing* in set theory – the *e pluribus unum* aspect of the theory – is a matter of taking mereological fusions of previously given items; so this part of set theory should not be a source of perplexity. The hypothesis about the size of reality is audacious (though it is perfectly *clear*, stated as it is in the pellucid language of pure mereology). But mathematics requires it, so it is warranted. This leaves the theory of the singleton function as the main potential locus of philosophical perplexity in the metaphysics of set theory.

To see why one might be puzzled, begin by noting that the official axioms are consistent with the supposition that singletons are parts of ordinary, spatiotemporal particulars. Some speck of matter in your left thumb might be your singleton, or the singleton of the Crab Nebula. But that would entail that you are not an individual, since you would then have a singleton as a part, and that would leave us with two possibilities. One is that all of your atomic parts are singletons, in which case you would be a class, perhaps a class containing the Crab Nebula.[11] The other is that you are a *mixed fusion* composed of singletons and individuals together. The first possibility is bizarre, but the second is totally unacceptable. It is a consequence of the *Domain* axiom that mixed fusions do not have singletons and are therefore ineligible for membership. But it is a settled fact – the basis for applications of set theory outside mathematics – that ordinary things can be members of sets. So: while this is not strictly forced upon us by the official axioms, we can say with confidence that the atoms that compose the sets – the singletons – are not ordinary spatiotemporal particulars: specks of matter, elementary particles, spacetime points, and so on. Whatever they are, they are . . . something else.[12]

We may note in passing that Lewis's insistence that mixed fusions do not have singletons has striking implications when combined with the central claim of his set theoretic constructionalism,

that is, that properties and relations are to be identified with classes. Consider the fusion of you and your singleton: you + {you}. This item has no singleton and is therefore ineligible for membership. So if properties and relations are classes, *this item has no properties and stands in no relations*. You might have thought that it would have the property of *being a mixed fusion*, or that that you would stand in the *part–whole relation* to it. But for Lewis there is no property had by all and only the mixed fusions, and no relation in which *A* stands to *B* iff *A* is part of *B*.[13] Now of course *any* view that identifies properties with classes will have annoying consequences of this sort. Since there is no class of all classes in most versions of set theory, there will be no property of being a class and no relation of class membership. The point worth stressing is that Lewis's theory has an especially acute version of this problem. There is nothing intrinsically paradoxical in the supposition of a part–whole relation. The relation is only problematic given Lewis's distinctive view of sets together with his set theoretic account of properties and relations.[14]

Returning to the main thread, we were stressing just how little we know about singletons, and hence about the classes that they compose, beyond the observation that singletons are 'extraordinary atoms' that are not parts of individuals. The axioms guarantee that some atom is your singleton. So where is it? What is it like intrinsically? What is it about this particular metaphysical speck that makes it *your* singleton rather than *mine*? These are meaningful questions on the Simple View, but as Lewis concedes, we have no idea how they are to be answered.

In *Plurality* Lewis suggests that a set is where its members are (Lewis 1986a, 94–5). This entails that your singleton is where you are, and hence that your singleton is an *extended simple*: an atom that occupies an extended region of spacetime.[15] Indeed the view entails that infinite stack of singletons based on you – {you}, {{you}}, etc. – is a collection of extended simples all of which occupy exactly the same region. This is not obviously wrong, but it is peculiar; Lewis 1991, 32).

This weird view would give us give us some purchase on the intrinsic nature of singletons. (The singleton of a long-lived St. Bernard would be bigger than the singleton of a short-lived Yorkie.) And it would yield a necessary condition for a thing to be *your* singleton: it must be an extended simple that is where you are. But beyond this it tells us nothing. We may know that {you} and {{you}} are co-located extended simples. But we have no idea whether they differ in other ways, and if asked why one of these suitably shaped items is *your* singleton rather than (say) your singleton's singleton, our spade is turned.

Lewis's considered view, however, is that even this much is uncertain:

> I don't say the classes are in space and time. I don't say they aren't. I say we're in the sad fix that we haven't a clue whether they are or whether they aren't. (Lewis 1991, 33)

Nor do we have any inkling as to the intrinsic qualities of these singletons. It is consistent with everything we know that they are featureless abstract points. But it is also consistent with what we know that they bear qualities, though of course we have no idea what those qualities might be.

This is not a decisive objection to the view. As Lewis says in another context, "Who ever promised me that I was capable in principle of knowing everything?" (2009, 211). The more serious worry, according to Lewis, is that we cannot answer certain elementary questions about the relation of member to singleton, and this threatens our claim to *understand* the language of set theory.[16] The most important such question poses what I shall call *the classification problem* for the Simple View: "Is [the member-singleton relation] an external relation, like the spatiotemporal relations of distance? Or is it an internal relation, like a relation of intrinsic similarity or difference? Or is it a combination of the two? Or something else altogether?" (Lewis 1991, 34). Lewis holds that any respectable theory should be able to classify its primitive relations (Lewis 1986a, 176–7). It's bad enough that the Simple

View cannot do this. The real problem, however, is not that the view is uninformative: it is that all of the more opinionated versions of it are problematic.

The relation of member to singleton cannot be internal, since two qualitatively identical individuals will have distinct singletons, whereas qualitatively identical things can never differ in the internal relations they bear. Moreover it seems clear that it cannot be an external relation of any familiar sort (e.g. spatiotemporal distance). Lewis leaves open the possibility that it might be an external relation of some utterly different sort (Lewis 1991, 37), but concedes that we have no inkling what this relation might be.

Lewis gives short shrift to the possibility the relation of member to singleton is neither internal nor external, but that view is worth taking seriously. Internal and external relations are *qualitative*. In either case, whether X bears R to Y supervenes on the qualitative intrinsic natures of certain things: of X and Y (in the case of internal relations) or of the mereological aggregate $X + Y$ (in the case of external relations). But consider a pair of perfect twins, Romulus and Remus, and their singletons. It is frankly nuts to suppose that the atom that is Romulus's singleton counts as *his* singleton in virtue of (say) its color or some other intrinsic quality, or in virtue of its distance from him, or anything of the sort. The only thing to think is that the fact that {Romulus} is the singleton of Romulus is not grounded in the *qualities* of these items at all. How can this be? Suppose for a moment that the extraordinary atoms that play the role of singletons are featureless abstract points that stand in no external relations to concrete things. If there are sufficiently many of these specks, there will be functions that map small objects onto abstract atoms in conformity with the axioms. Each such function is a non-qualitative relation, neither internal nor external. Lewis cannot deny the *existence* of non-qualitative relations.[17] And since it is mad to suppose that color-like intrinsic properties and distance-like external relations play a role in determining which atom counts as your singleton, the proponent of the Simple View should regard it as a serious possibility that the relation of member to singleton is a non-qualitative relation of this sort.[18]

In any case, the core metaphysical commitment of the Simple View may be put as follows: every small item in reality (mixed fusions aside) is associated with what might as well be its very own featureless abstract speck. *Classes are mereological aggregates of these specks*. The only relevant intrinsic property of a class is the number of specks that make it up. If there are just as many dogs as cats, then the class of dogs and the class of cats might well be intrinsically just alike. If we ask why Fido belongs to one but not the other, we can say: he belongs to the class of dogs because his singleton is part of the class of dogs. But if we ask why a certain abstract speck counts as *Fido's* singleton, our spade is turned.

24.3 Metaphysics of Classes: The Structuralist View

The Structuralist View of classes is designed to mitigate the weirdness of the Simple View. Note first that if there is one function from small things to atoms that satisfies the axioms for 'singleton', then there are many. In particular, if s satisfies the axioms, then so does every permutation of s, e.g. s^*:

$$s^*(x) = s(\text{you}) \text{ if } x = \text{me}$$
$$\quad\quad s(\text{me}) \text{ if } x = \text{you}$$
$$\quad\quad s(x) \text{ otherwise}$$

The Simple View assumes that, of the many functions that satisfy the axioms, the word 'singleton' somehow picks out exactly one of them. This is what renders the view vulnerable to the awkward

questions. For only given this assumption does it makes sense to ask, "What is it about {you} that makes this item your singleton rather than mine?"

On the Structuralist View, the various singleton functions are on a par. There is no such thing as *the* singleton function, hence no such thing as *your* singleton. Rather you have one singleton relative to *s*, and a different singleton relative to *s**. Claims about classes are to be understood, not as claims about a single determinate mapping from small things to extraordinary atoms, but rather as general claims about what all of the singleton functions have in common.

There are two ways to implement this proposal. (Lewis mentions both and treats them as equivalent, but they are different, as we shall see.) Let Σ(singleton) be the conjunction of the axioms governing the singleton function[19], and let Φ(singleton) be any claim about classes, written out in primitive notation with 'singleton' as the only set-theoretic expression, as Lewis has shown us how to do. Then according to the *quantificational variant* of the Structuralist View, the real content of Φ(singleton) is given by the conjunction:

> There is a function *s* such that $\Sigma(s)$, and
> For any *s*, if $\Sigma(s)$ then $\Phi(s)$.

Since the important truth about sets (including the axioms and theorems of set theory itself) will be true relative to *any* singleton function, this proposal is non-revisionary. But now let *a* be an atom eligible to serve as a singleton, and consider the claim

> $a = \{\text{you}\}$

Some singleton functions assign *a* to you as your singleton; others do not. And from this it follows that the claim '$a = \{\text{you}\}$' is *false* on the quantificational variant of the Structuralist View. But of course *a* was chosen arbitrarily. So for each atom *x*, *x* is not your singleton. In one respect, this is a good thing, since it explains why we are under no pressure to say what it is about your singleton that makes it yours. But in another respect, it is awkward in a familiar way. After all we know that you have a unique singleton. This is true on the structuralist variant because for every singleton function *s*, $\exists!x\ x = s(\text{you})$. But we also know that *a* is not your singleton, and *b* is not your singleton, . . . where *a*, *b*, etc., are all the atoms that exist. This can sound contradictory unless we are very careful to remember what we mean by the claim that you have a (unique) singleton.

According to the *supervaluational variant*, the claims of set theory do not involve hidden quantification over singleton functions. Rather we think of the word 'singleton' as *semantically indeterminate*, as *dividing its reference*[20] over each of the candidate meanings it might have had if we had bothered (*per impossibile*) to fix on a single meaning. "Think of it this way," Lewis says:

> The word 'singleton' is highly equivocal, since all of the interpretations that satisfy the conjoined axioms are equally 'intended' interpretations. Though it appears in the guise of a primitive constant, it is no better than a variable – albeit a variable restricted to a certain limited range of admissible values. The equivocal sentence about [you and your singleton] is true *simpliciter* iff it is true on all its intended interpretations. (Lewis 1991, 47)

And we might add, as is customary in contexts of this sort: the equivocal sentence is false *simpliciter* iff it is false on all its intended interpretations; otherwise it is neither true nor false.[21] On this account, the sentence '$a = \{\text{you}\}$' is neither true nor false, as is its negation. So we cannot say that *a* is *not* your singleton. What we can say instead is that *a* is not *determinately* your singleton.

The differences between these two views are subtle. Where one sees hidden quantification, the other sees semantic indecision. The result is that certain claims that must be reckoned false on the quantificational approach are rendered truth-valueless on the supervaluational approach. But this small difference is potentially consequential, as we shall see.[22]

Whether understood in the quantificational way or in the supervaluational way, the main advantage of the Structuralist View, according to Lewis, is that it evades the demand that we say what it is about *your singleton* that makes it yours, and also the demand that we classify the singleton relation as internal, external or neither. These demands are evaded because on the Structuralist View, there is no such relation as *the* singleton relation, and hence no such (determinate) item as *your* singleton.

It may be wondered, however, what sort of advantage this is. The Simple View and Structuralist View posit the same array of singleton relations, most of which are non-qualitative. According to the Simple View, our word 'singleton' somehow latches onto one of them; the Structuralist View denies this. But as we have noted, Lewis himself provides a mechanism for securing this sort of semantic determinacy (Lewis 1983).[23] The alleged advantage of structuralism thus comes down to this: it posits *one less application of the reference fixing mechanism*. It is unclear whether this should be reckoned a significant advantage.

On the other hand, the Structuralist View has two distinctive problems of its own. First, it requires the theorist to make sense of quantification over relations – or better: quantification into sentential positions normally occupied by binary relation symbols – without assuming set theory. The main technical achievement of the appendix to *Parts of Classes* is to solve this problem.[24]

Second, Lewis worries that as a philosophy of mathematics, set theoretic structuralism is objectionably innovative:

> Unfortunately, the structuralist redemption of set theory is not retroactive. Structuralist set theory really is new. It would be high-handed anachronism to claim that set theory was structuralist all along . . . Before mid-1989, nobody even knew how to be a set theoretic structuralist. To be a structuralist, you quantify over relations; to be a set theoretic structuralist, you do so before you are entitled to the resources of set theory; and that is what we've only now learned how to do. (Lewis 1991, 53)

In the philosophy of mathematics, Lewis is a principled conservative. The philosopher's job is not revise mathematics on philosophical grounds, but to make sense of mathematics as it is (Lewis 1991). Since the Structuralist View requires new technical resources, it is new. Whatever its merits, it is not an account of set theory as we find it.

This is another point at which it is useful to distinguish the two versions of structuralism outlined above. The quantificational variant is clearly new in the way Lewis describes, but the supervaluational variant is not. The language of set theory on the supervaluational view is just the language of standard set theory. The view posits no hidden quantifiers in the mathematical theory itself (though it does make use of quantifiers over relations in giving a semantic theory for that language). It is perfectly common for people to use a word that is indeterminate in reference without realizing that they are doing so, and without having the resources to state a supervaluational theory of truth for their language. Before 1966, no one knew how to do *that*! But semantic indecision was rife from the dawn of language. If mathematicians were firmly committed to the view that the language of set theory is *not* rife with indeterminacy, then this view would be at odds with their self-understanding. But (a) there is no reason to suppose that mathematicians are committed to the semantic determinacy of the language of set theory, and (b) even if they were, the amateur semantic theorizing of mathematicians has nothing like the authority we accord their mathematical opinions.

In any case, the metaphysical content of the Structuralist View may be put as follows. Set theory requires the existence of inaccessibly many extraordinary simples. But there is no such thing as *the* singleton function, hence no such thing as *the* class of *F*s. Terms that purport to denote classes are either wildly indeterminate in reference (as on the supervaluational view) or they are not really closed terms at all. In either case, weighty metaphysical questions about the natures of sets do not arise.

24.4 Implications of the Simple View

We turn now to the implications of Lewis's account of classes for his larger ontological scheme, beginning with the Simple View. The main feature of this view, to repeat, is that classes are aggregates of what may as well be featureless abstract specks, and the relation of member to singleton is a hard to classify primitive relation known only by its formal properties. Suppose this is right, and suppose further that philosophical entities like propositions, properties, meanings, types, and events are all identified with set theoretic constructions from concrete possibilia. It then follows that *these items are likewise fusions of what may as well be featureless abstract points*.

Consider the colors *red*, *orange*, and *blue*. For Lewis these are the fusions of the singletons of the red, orange, and blue things respectively. Since there are presumably just as many red things as there are orange things and blue things, these properties will have the same number of atomic parts. And since these parts may as well be featureless abstract whatnots, it follows that, for all we know, the colors are all intrinsically just alike.

This is a surprise. It is often said that *red* resembles *orange* more closely than it resembles *blue* (Jackson 1977). Lewis's view entails that if this claim is taken at face value, we have no reason to accept it. Of course it is open to Lewis to paraphrase the similarity claim as a claim, not about the intrinsic natures of colors, but rather about the resemblance of particulars.[25] It is also open to Lewis to maintain that insofar as the claim of similarity is true, it does not concern properties, but rather (say) Armstrong-style universals. (This is awkward in this case, since the visible colors are presumably not genuine universals in Armstrong's sense.[26]) What Lewis cannot say, with any confidence, is that the claim is true when taken as a claim about the items he calls properties.

Similar problems will arise whenever philosophical entities appear to exhibit intrinsic similarities and differences. You might have thought that the English word 'dog' is more like the word 'dot' than it is like the Basque word 'urtxintxa'. But if a word type is identified with the class of its tokens (or with some more complex set theoretic construction from phonemes and other linguistic features) then for all we know these words are all intrinsically indiscernible aggregates of abstract points, in which case the similarity claim is false if taken literally.

An especially dramatic version of the problem concerns events. Suppose we have an iron sphere *S* that is simultaneously spinning and heating up. Following Davidson (1980, 178), Lewis holds that the spinning and the warming are distinct events even though they occupy the same region of spacetime. This counts against the Quinean identification of events with spacetime regions, favoring instead the identification of events with certain *properties* – that is, classes – of spacetime regions. More specifically, Lewis identifies the spinning with a class of regions, at most one per world, in which a counterpart of *S* spins more or less as *S* actually does, whereas the warming is identified with a different class: the class of the regions in which *S* gets warmer (Lewis 1986b). But now consider three events: Game 6 of the 1986 World Series, Game 7 of the 1986 World Series, and the Big Bang. If events are classes, and classes are sums of what may as well be featureless abstract points, then we cannot say that the first two are more similar to one another than either is to the third. Of course we can say that the spatiotemporal regions in which the baseball games take place are more similar,

but that is not a claim about the intrinsic natures of the events in question. The same cluster of doctrines has the equally striking implication that when we *see* an event we learn nothing whatsoever about its intrinsic nature, since we clearly learn nothing about the intrinsic nature of the sum of extraordinary atoms that is the event on Lewis's account.

Putting these observations together, we may draw the obvious conclusion. When Lewis's set theoretic reductionism about the philosophical entities is married to the Simple View of classes, the result implies that for all we know, the property of being red, the 1986 World Series, and the word 'urtxintxa' are intrinsically just alike. Nothing about the natures of these items (beyond their cardinality) fits them for the roles they play. If a certain item is the World Series, and not the word 'urtxintxa', that is *entirely* thanks to the relations in which it stands to other things.

Our complete ignorance of the intrinsic natures of properties and other philosophical entities on the Simple View has one more noteworthy implication. As we have noted, one central ingredient in Lewis's mature philosophy is the idea that some properties are more *natural* than others, and that some are to be reckoned *perfectly* natural. Lewis allows that these notions might be explained in other term: the perfectly natural properties might be identified with Armstrong-style immanent universals. But he inclines to the view that that naturalness is a primitive feature of the items he calls properties, with certain classes of individuals singled out as especially eligible to serve as the referents of predicates, or to figure in the laws of nature (Lewis 1983). Positive unit charge might be a perfectly natural property, and even if it is not perfectly natural, it is certainly more natural than (say) *banana-charge* – the property of being either positively charged or a banana. Considered as classes, however, these items might be intrinsically just alike. It follows that *naturalness is not an intrinsic property of properties, and relative naturalness is not an internal relation*. This is moderately surprising. Lewis's inegalitarianism is naturally taken as the view that something in the *nature* of unit positive charge renders it especially natural, hence especially eligible to serve as the referent of a predicate, etc. But Lewis cannot say this, given his view that properties are classes together with his view that classes are composed of what may as well be featureless abstract specks.

If naturalness is not a qualitative, intrinsic property of properties, what sort of property is it? It is open to Lewis to say that naturalness is a non-relational, non-qualitative property of classes, one possessed by the class of positively charged items but not by the class of banana-charged items, even though these classes may be qualitatively just alike. Lewis's best option, however, is to say that the naturalness of a class derives, not from any intrinsic feature of the class itself, but rather from a feature of its elements. Unit positive charge may count as perfectly natural, for example, because (a) it is the class that contains the positively charged things, and (b) the positively charged things are *collectively perfectly natural*. Similarly, positive charge might be more natural than banana charge because the positively charged things are (collectively) more natural than the banana-charged things. On this account, the core ideological commitment of the inegalitarian view of properties is not a special primitive of properties, but rather an irreducibly plural predicate of individuals. The naturalness of *properties* is grounded in the naturalness of *things*.

24.5 Implications of the Structuralist View

On the Simple View, there is such a thing as the class of red things, but we know next to nothing about its nature. On the Structuralist View, by contrast, *there is no such thing as the class of red things* (to put the point tendentiously), and hence, given Lewis's account of the philosophical entities, no such thing as the color red, or the proposition that snow is white, or the Big Bang. This formulation is tendentious, of course, since the sentence "There is such a thing as the class of red

things" is true relative to every singleton function, hence true *simpliciter* according to the structuralist. And yet it remains the case that if we could survey absolutely everything, we would never come across a thing of which we could truly say: *that* is the class of red things. Instead, absolutely every aggregate composed of a suitable number of suitable atoms will count as the class of red things relative to some singleton function, and as some other class, or no class at all, relative to others. And that means that every suitable aggregate will count as *the color red* relative to some singleton functions and as some other property, or perhaps as a baseball game or as the word 'platypus', relative to others. On the supervaluationist variant, the point can be put as follows: every linguistic expression that purports to pick out a property, relation, event, proposition, etc., is in fact *wildly* indeterminate in reference. On the quantificational variant such expressions are not really devices of reference at all: they are open expression containing a free variable bound by a quantifier over singleton functions.

Suppose we witness a battle and say, "Let's call that one 'the Battle of the Bulge,' " or suppose we're learning a foreign alphabet and someone tells us, "That letter is called 'aleph'." We are accustomed to the idea that terms of this sort may exhibit moderate semantic indeterminacy. If it is unclear exactly when the battle begins or ends, or which squiggles in the history of writing count as tokens of the letter aleph, then there will be many equally good candidate referents for these terms, just as there are many equally good candidate referents for 'Mt. Blanc'. But there is indeterminacy and then there is indeterminacy. It is one thing to say that it is unsettled which of certain massively overlapping hunks of rock and snow counts as Mt. Blanc. It is another thing to say, with Quine (Quine 1969) that every term divides its reference over *just about every object in the universe*. Lewis's metasemantic account in Lewis (1983 and 1984) is designed in large part to block this sort of radical indeterminacy. But even if that account is successful in the case of terms for individuals, it has no application when it comes to names for classes on the Structuralist View. When the Structuralist View is married to Lewis's identification of philosophical entities with classes, the result is widespread and radical indeterminacy of reference. For any term t that is not a name for an individual and any aggregate o of extraordinary atoms of the appropriate cardinality, there is an admissible interpretation of the language relative to which t refers to o.

Again, this is surprising. We are normally inclined to suppose that just as we often have a single, determinate rabbit in mind when we think of (or point at) Flopsy, so we often have a single, determinate species – that is, a single determinate class – in mind when we speak of the domestic rabbit, *Oryctolagus cuniculus* (modulo inevitable fuzziness at the boundaries), or of the property of being a rabbit. This is clearest when it comes to terms for sensible qualities, observable types, and other philosophical entities that seem to us to be *given* in experience. When we see a round ball and focus first on its color and then on its shape, it is hard to resist the sense that we have shifted our attention from one relatively determinate property to another. If the Structuralist View is correct, this is an illusion.

Turn now to the theory of natural properties. In our previous discussion we asked whether the naturalness of a property is best understood as intrinsic or extrinsic, and we noted that while it was open to Lewis to say that naturalness was a non-qualitative, non-relational property of classes, it was more plausible to treat it as an extrinsic feature possessed by a class when its elements are *collectively* natural. On the structuralist account, the first option is foreclosed. Since the expressions 'positive unit charge' and 'banana charge' are almost maximally indeterminate in reference, it makes no sense to suppose that one but not the other picks out a class that has the non-qualitative property of perfect naturalness. On the Structuralist View, any aggregate of abstract whatnots of a suitable cardinality will count as the property of positive charge – hence as natural – relative to some singleton functions but not others. So if 'positive charge is a natural property' comes out true relative to every singleton

function, this will not be thanks to how things stand with the various referents of 'positive charge'. Rather it will be so because the positively charged things are collectively natural, from which it follows that for any singleton function *s*, the class of positively charged things *relative to s* is a natural class *relative to s*. This version of the view thus makes it even more vivid than the last that when Lewis's identification of properties is married to his distinctive account of classes, the *naturalness of classes* does no explanatory work. The work is done entirely by the naturalness of certain pluralities of individuals.

24.6 The Case against Magical Ersatzism Revisited

This section explores the implications of Lewis's set theoretic constructionalism for one of the most widely discussed arguments in *On the Plurality of Worlds*: the case against magical ersatzism (Lewis 1986a, 176–87). Magical ersatzism is an alternative to modal realism according to which possible worlds are not concrete universes rather like "us and our surroundings," but rather abstract entities with no relevant internal structure. Anyone who speaks the language of possible worlds faces a question: What makes any given world a world at which – or *according to which* – a donkey talks? The modal realist can say: the worlds according to which a donkey talks are precisely the worlds that have a talking donkey as a part. They have this "representational property" in virtue of their intrinsic make up. The magical ersatzer cannot say this. What can he say?

Lewis imagines an ersatzer who responds by presenting a little theory (paraphrased below).

> The abstract simples we are calling worlds are a subset of a larger group of abstract simples that we call *elements*. (Others call them *propositions*.) Some of these elements are *selected* by the actual concrete universe. Which elements are selected depends on how things are. If a donkey talks, some elements are selected; if a pig flies, others are selected. Selection is a primitive relation, but given selection, we can define a relation of implication on the elements: *p* implies *q* iff necessarily, if the actual world selects *p*, it selects *q*. And then we can tell you which of these abstract simples are worlds. The elements form a complete Boolean algebra under implication, and the worlds are the maximal elements in that algebra. They are the elements that imply, but are not implied by, other elements (except the 'null' element that implies all of the elements). *w* is a world *at which* donkeys talk iff (it is selected iff donkeys talk).

Lewis does not object that the ersatzer must take selection (or as it is more commonly called, *truth*) as primitive. But he does think that this choice places the ersatzer in an awkward predicament. Even if he cannot define selection, he should be able to classify it. We know that it's a relation between a concrete thing – the concrete universe – and certain abstract objects. Lewis wants to know whether it is an internal relation, an external relation, or neither, and he thinks that the ersatzer cannot give a decent answer.

An internal relation supervenes on the intrinsic characters of the relata. So if selection is an internal relation, the simple that is selected iff a donkey talks must differ qualitatively from the simple that is selected iff a pig flies. But since they are abstract, we have no inkling what the relevant qualitative features might be, or why the possession of one of them – a lovely pink, say – should guarantee that a simple is selected iff a donkey talks. An external relation is like a distance relation: it supervenes on the intrinsic character of the relata *taken together*. But it is just nuts to suppose that the abstract 'elements' are situated at something like a distance from the concrete universe, with some of them – the selected ones – hovering nearby, while rest hover at some greater distance. So it is hard to believe that selection is internal, and equally hard to believe that it is external.[27]

The best option for the ersatzer is to say that the relation is non-qualitative, and therefore neither internal nor external. To take this route is abandon all hope of saying what it is about *this* abstract simple in virtue of which it is selected iff a donkey talks. Suppose we had a red ball and 100 non-spatial, qualitatively indiscernible abstract points that stand in no relevant external relation to the ball. Now consider a relation in which this ball stands to 99 of those simples, but not the 100th. There certainly is such a relation, considered as a class of ordered pairs, and a moment's reflection will show that it is neither internal nor external. The ersatzer's relation of selection must be like *that*.

Lewis regards this view as untenable, but before we turn to his objection, let us see how Lewis solves the corresponding problem for modal realism. For Lewis, the items that play the role that elements play in the ersatzer's theory are worlds and sets of worlds. When it comes to worlds, the account is trivial: the selection relation is just the relation of exact qualitative similarity: each world selects itself and any perfect qualitative duplicate thereof, so the selection relation is internal. But now consider propositions, bearing in mind that for Lewis, a proposition is a class of worlds.[28] A world *selects* a proposition just in case it is a member of it. So for the modal realist, *the problem of classifying the selection relation reduces to the problem of classifying the membership relation* (or more specifically, the membership relation as restricted to worlds and classes of worlds).

How is this relation to be classified? On the Simple View, these classes may as well be aggregates of featureless whatnots, so the relation cannot be internal. (For all we know, the class of worlds at which a donkey talks is intrinsically identical to the class of worlds at which a donkey brays. But the actual world selects only the latter.) Moreover it beggars imagination to suppose that it is external in the manner of a distance relation. So unless Lewis can somehow salvage this bizarre option, *the modal realist will wind up saying exactly what the magical ersatzer says* in response to this classification problem: insofar as it concerns propositions, the relation of selection is neither internal nor external.

On the Structuralist View, the problem may not arise. Since there is no item that is, determinately, the class of worlds at which a donkey talks, it makes no clear sense to ask what it is about *this item* that makes it the proposition that is selected iff a donkey talks. A choice of singleton function induces a selection relation. But since there is no such thing as *the* singleton function, there is no such thing as *the* selection relation. And yet it remains the case that many of these selection relations are non-qualitative. Indeed, if the abstract atoms from which the classes are constructed are featureless and stand in no external relations to the concrete worlds, it will follow that *all* of the many selection relations are non-qualitative, in which case the modal realist will again be forced to say exactly what the ersatzer says: (relative to every singleton relation), the selection relation is neither internal nor external.

The upshot is that Lewis's classification argument against magical ersatzism is inconclusive. Whether propositions are construed as abstract simples *or as classes of concrete worlds*, the relation of selection is almost certainly non-qualitative. So in the end, there is no clear advantage for the realist.[29]

With this in mind, it may be useful to review Lewis's brief comment on the suggestion that the ersatzer might treat selection as a non-qualitative relation – a suggestion he clearly views as desperate. The only version of the view Lewis considers holds that

> Selection is not any kind of natural relation. There is nothing to it but a list – a list entirely out of reach of thought and language. Element E is selected iff:
>
> E is N_1 and P_1; or E is N_2 and P_2; or…
>
> where the Ns are mere proper names of elements. No such names could possibly be available to us; so again it's magic if the ersatzer understands his own primitive. (Lewis 1986a, 182)

The key claim is the suggestion that a non-qualitative relation cannot be a relatively natural relation. This makes perfect sense if we suppose that the natural properties and relations are those that "make for qualitative similarity" among their instances (Lewis 1986a, 60). But if we say instead that the natural properties and relations are those that are especially eligible to serve as referents for words or to figure in the contents of our thoughts, then it is not obvious that a non-qualitative property/relation cannot be natural.[30] But if this is possible, then the ersatzer can understand his primitive simply by understanding the rules governing its use, leaving it up to 'nature' to see to it that the expression governed by those rules attaches to the most natural relation in the vicinity. We have suggested that the proponent of the Simple View of classes should take this view very seriously, treating the singleton relation as a natural, especially eligible, non-qualitative relation. If this view is available in that context, it is available to the magical ersatzer as well.

It remains to note one possibility which we will not have space to explore. The proponent of the Simple View of classes and the proponent of a magical ersatzist account of worlds face analogous challenges: to classify the relations they take as primitive as either internal, external or neither. In his discussion of classes, Lewis seeks to evade the challenge by going structuralist. Can the magical ersatzer do likewise? The ersatzer's theory of worlds and propositions is cast in terms of what seems to be a primitive relation of selection that holds contingently between the single concrete world and certain abstract objects. But suppose the ersatzer replies as follows:

> The word 'selects' in my theory may be highly indeterminate in reference. If there is one especially natural relation between the concrete universe and abstract entities that makes sense of what I say, so be it. But if not, the word divides its reference over all of the equally natural relations that can do the job. What's true/false in my language of worlds and propositions is what's true/false on every admissible interpretation of this language. This means that lots of apparently meaningful questions about worlds and propositions will not have answers. In particular, if you ask me to focus on the element (i.e. proposition) that is selected if and only if donkeys talk and then to say why *this* abstract simple has this impressive feature, I will reject your question. It's a bit like asking why a certain determinate hunk of rock and snow counts as Mt. Blanc. The question is misguided because it presupposes that a single determinate item has somehow beaten out its close cousins in the competition to play a certain role. If no determinate item has been cast in the role, then it makes no sense to ask why *it* has won the part.

This sort of structuralism about worlds and propositions has much to recommend it. It evades Lewis's challenges to the magical ersatzist without invoking a primitive non-qualitative natural relation of selection, and it comports with the widespread sense among linguists, philosophers, and others who make casual use of the idiom of worlds and propositions that Lewis's metaphysical seriousness about these idioms is somehow misguided. It renders the metaphysical obligations of these theories so light as to be almost insubstantial: all we require (in effect) is sufficiently many abstract entities of whatever sort: their natures make no difference whatsoever.

The main difficulty for this proposal is to make sense of structuralism in this context. In stating his proposal, the structuralist must at some point quantify over the 'admissible' semantic values of the expression he regards as indeterminate in reference. In the case of Lewis's set theory, this amounts to quantification over relations-in-extension, and it takes a great deal of careful work to show what this can mean in a context in which set theory is not yet available. The would-be structuralist about worlds and propositions faces a different challenge. The theory in terms of which he introduces his key notions is a modal theory. It will include (either as theorems or as axioms) statements like

> There is an element *e* such that necessarily, *e* is selected iff a donkey talks.

The structuralist's claim is that 'element' and 'selected' are indeterminate in reference, or alternatively, that they amount to variables bound by a hidden quantifier. But how shall he characterize the items over which they divide their reference, or over which these quantifiers range? We know that the elements are supposed to be abstract simples. So relative to any admissible interpretation of the language, 'element' will pick out some determinate class of abstract simples. But what sort of item will 'select' pick out? A relation, you say. But what sort of thing is that? Not a class of ordered pairs. For it is supposed to be a *contingent* matter whether the concrete world α selects e; but it is not a contingent matter that the pair $<\alpha, e>$ belongs to some class of ordered pairs. No; if the view is to make sense, an admissible interpretation of the language must associate 'select' with a relation-in-intension. If we had possible worlds to work with, we could take these to be functions from worlds to classes of pairs. But just as Lewis cannot take sets for granted in explaining his set theoretic structuralism, so the would-be structuralist about worlds and propositions cannot take these items for granted in setting out his view. We leave it open whether this problem can be solved.

24.7 Conclusion

According to Lewis's set theoretic constructionalism, the theoretical entities of philosophy – propositions, properties, relations, meanings, events, etc. – are set theoretic constructions from possibilia. According to Lewis's mature account of sets, a set is an aggregate of 'extraordinary atoms' with no relevant intrinsic features. (They may have intrinsic features, but these features play no role in the account.) When these two aspects of Lewis's thought are brought together, the results are surprising. On the Simple View of classes, the view entails that for all we know, the philosophical entities are all intrinsically just alike (unless they happen to differ in cardinality), and hence that for all we know, there is no intrinsic difference between the color purple and the 1986 World Series. On the Structuralist View, terms that purport refer to philosophical entities are wildly indeterminate in reference. These consequences do not refute Lewis's view. It is a safe bet that *any* worked out account of the philosophical entities will have strange consequences. But since Lewis's case for his distinctive mix of modal realism and set theoretic constructionalism ultimately rests on an assessment of the costs and benefits of the theory as a whole, it is important to know what the view implies. Only then can we say with any confidence that it's worth the price.

Notes

1 Lewis entertains the possibility of spatially located universals, but such items play no role in his positive account. Mereological mixtures of sets and individuals exist for anyone who accepts both set theory and unrestricted mereological composition, but they are junk and Lewis makes no use of them.

2 Set theoretic constructionalism begins in the foundations of mathematics, and moves into general philosophy with Russell's famous remark that "the supreme maxim in scientific theorizing is this: Wherever possible, logical constructions are to be substituted for inferred entities" (Russell 1988, 115), cited by Carnap as the epigraph to Carnap 1969 (5). Quine and Lewis deviate from these writers in two relevant ways. As regards the *basis*, both Russell and Carnap regard the basic individuals, not as concrete spatiotemporal entities, but as 'phenomenal' items akin to sense data. Second, as regards the apparatus, both Russell and Carnap (in the relevant texts) regard classes as *logical fictions*, whereas Quine and Lewis show no appetite for this doctrine. Carnap's version of set theoretic constructionalism is analyzed and rejected in Goodman 1977. For an articulation of set theoretic constructionalism that was especially important for Lewis, see Montague 1969.

3 An exception is the discussion at Lewis 1986a (94–5), which pre-dates the mature theory of classes that is our focus here.
4 A point emphasized by Goodman 1979.
5 This assumes that there is only one object composed of *a* and *b* alone, a special case of the thesis of *uniqueness* of composition. For Lewis's defense of this and other principles of classical mereology, see Lewis 1991, §3.4–5.
6 The parenthetical restriction to *non-empty* subclasses is redundant. Lewis acknowledges the empty *set* – which he identifies, idiosyncratically, with the fusion of all of the individuals in reality. But he reserves the word *class* for things with members. So every class is non-empty, by definition.
7 This is a stipulation that serves Lewis's purposes. It is not meant as a profound discovery about the nature of "that than which none emptier can be conceived" (van Frassen 1975, 39).
8 The hypothesis asserts that there are *inaccessibly many atoms*. Despite its apparently mathematical character, Lewis shows that it can be stated in the language of mereology supplemented with monadic plural quantifiers (Lewis 1991, 93–5).
9 A fusion of singletons is *small* iff it contains fewer atomic parts than there are atoms in reality.
10 See Burgess (chapter 29, this volume) for discussion.
11 We can rule out the possibility that you are a class and some material speck in your left thumb is *your* singleton. For then you would be a self-membered class, and that's not allowed.
12 Lewis does not draw this conclusion, but he does draw a limited version of it. Given his view that is a relatively small (from the standpoint of set theory) upper bound to the number of ordinary spatiotemporal particulars in all of reality, Lewis concludes that *most* of the singletons are not ordinary spatiotemporal things (1986a, 89).
13 This has implications for semantics: in general, for any predicate *F* that is truly predicable of some mixed fusion, the meaning (semantic value) of *F* cannot be a class that contains the items of which *F* is truly predicable. So either such predicates do not have meanings, or their meanings are not classes of this sort.
14 Actually, the standard part–whole relation *is* problematic on almost any set theoretic account. The standard part–whole relation is unrestrictedly reflexive. But, in standard set theory, there are no unrestrictedly reflexive relations, considered as classes of pairs. (If there were, there would be a class of all classes, which there isn't.) The set theoretic reductionist can avoid this result by holding that classes are never parts – thereby restricting mereological composition in the least problematic way. Alternatively, he can say that even if there is no unrestricted part–whole relation, there is such a thing as the *proper–part* relation. But in order to say this he must maintain that classes have no proper parts. Lewis is committed to rejecting both approaches, and so to rejecting mereological relations.
15 On extended simples, see Hudson 2007.
16 Lewis repeats this worry in several places, e.g.: "It's a nasty predicament to claim that you somehow understand a primitive notion, although you have no idea how you could possibly understand it. That's the predicament I'm in when I accept the notion of singleton" (1991, 36), echoing Lewis 1986a (182). But the nature of the worry is somewhat unclear. The proponent of the Simple View maintains, in effect, that the singleton relation is a relation – perhaps *the unique natural relation* – that satisfies the official and unofficial axioms of set theory. Why isn't this knowledge sufficient to constitute *understanding* of the primitive? True, it does not put us in a position to answer certain questions about the relation. But so why should our ignorance of these facts threaten our claim to understand the word?
17 Take any qualitative relation, say, the relation of exact similarity in shape. Think of it as a class of ordered pairs. Now consider the relation that differs from this one by the deletion of some arbitrary pair, say <Romulus, Remus>. Or take an arbitrary mapping from objects in one possible world onto objects in another. The result is a relation that is neither internal nor external.
18 It may be asked how the word 'singleton' could possibly come to stand for one of these non-qualitative relations rather than another. But Lewis himself has the beginnings of an answer: If the word does refer determinately to one such relation, that is because on singleton relation is more *natural* than any of the others, where *naturalness* is a primitive property of relations. See Lewis 1983.

19 These include the mathematical axioms given above, together with a philosophical axiom to the effect that ordinary things like you and I count as individuals, and so have no singletons as parts. It is somewhat unclear how this axiom is to be formulated, but it is indispensable.
20 Field 1973.
21 This follows van Fraassen 1966, cited by Lewis in the footnote to the quoted passage.
22 Suppose that some of the candidate singletons are in space while others aren't, and suppose that there are sufficiently many candidates of both kinds to yield interpretations of 'singleton' on which all singletons are in space, and interpretations on which no singletons are in space. Then the quantificational variant will entail that 'All singletons are spatial' and 'Some singletons are non-spatial' are both *false*, whereas the supervaluational interpretation will entail that these claims are both truth-valueless.
23 This assumes that the *relatively natural* relations can include non-qualitative relations. Lewis does not discuss this possibility (though he is committed to regarding the *perfectly natural* properties as qualitative). For reason to think that he might have been amenable to this possibility, see the discussion of a mathematical example at Lewis 1983 (53).
24 For exposition, see Burgess, chapter 29, this volume.
25 Lewis entertains this possibility, notes its complexities, and does not adopt it (1983, 14–15).
26 As Lewis notes (1983, 17).
27 Lewis has a further objection to the idea that selection might be external. The hypothesis violates a plausible version of Hume's prohibition of necessary connections between distinct states of affairs. See Lewis 1991, 47.
28 Better: coarse-grained propositions are sets of worlds. Lewis allows that fine-grained propositions or structured meanings may also deserve the name. (Lewis 1986a, 57–9).
29 Peter van Inwagen famously argues that Lewis's complaints against the magical ersatz theory of propositions apply equally to his own theory of classes (Lewis 1986a, 180–1). The argument in this section aims to show that this argument does not just redound against Lewis's philosophy of mathematics: it redounds against Lewis's theory of *propositions* as well. For Lewis's response to van Inwagen, see van Inwagen 1986.
30 For some evidence that Lewis sometimes thinks of naturalness in these terms, consider his remarks on Kripkenstein's puzzle about 'quaddition' (Lewis 1983, 53). There Lewis suggests that the *addition* function must be a more natural relation than the *quaddition* function; but it is hard to see how this could be so if these pure mathematical relations must be qualitative.

References

Carnap, R.(1969). *The logical structure of the world and pseudoproblems in philosophy*. Trans. R.A. George. Chicago: Open Court.
Davidson, D. (1980). *Essays on actions and events*. New York: Oxford University Press.
Field, H. (1973). Theory change and the indeterminacy of reference. *Journal of Philosophy*, 70(14), 462–81.
Goodman, N. (1977). *The structure of appearance*, 3rd ed. Dordrecht: Springer.
Goodman, N. (1979). *Problems and projects*. Indianapolis: Hackett.
Hudson, H. (2007). Simples and gunk. *Philosophy Compass*, 2(2), 291–302.
Jackson, F. (1977). Statements about universals. *Mind*, 86(343), 427–9.
Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, 61(4), 343–77.
Lewis, D. (1984). Putnam's paradox. *Australasian Journal of Philosophy*, 62(3), 221–36.
Lewis, D. (1986a). *On the plurality of worlds*. Oxford: Blackwell.
Lewis, D. (1986b). Events. In D. Lewis, *Philosophical papers*, Vol. II, pp. 241–69. Oxford: Oxford University Press.
Lewis, D. (1991). *Parts of classes*. Oxford: Blackwell.
Lewis, D. (1993). Mathematics is megethology. *Philosophia Mathematica*, 1(1), 3–23.
Lewis, D. (2009). Ramseyan humility. In D. Braddon-Mitchell and R. Nola (Eds.), *Conceptual analysis and philosophical naturalism*, pp. 203–22. Cambridge, MA: MIT Press.

Montague, R. (1969). On the nature of certain philosophical entities. *The Monist*, 53(2), 159–94.
Quine, W.V. (1969). *Ontological relativity, and other essays*. New York: Columbia University Press.
Russell, B. (1988). *Mysticism and logic*. London: Rowman & Littlefield.
van Fraassen, B. (1966). Singular terms, truth-value gaps, and free logic. *Journal of Philosophy*, 63(17), 481–95.
van Fraassen, B. (1975). Platonism's pyrrhic victory. In A.L. Anderson, R.B. Marcus, R.M. Martin & F.B. Fitch (Eds.), *The logical enterprise*, pp. 39–50. New Haven, CT: Yale University Press.
van Inwagen, P. (1986). Two concepts of possible worlds. *Midwest Studies in Philosophy*, 11(1), 185–213.

25

Primitive Self-Ascription

Lewis on the De Se

RICHARD HOLTON

In a memorable children's story, Winnie-the-Pooh follows the tracks of what he thinks might be a woozle, until he realizes that he has been "Foolish and Deluded" and that the tracks are his own.[1] Similar ideas recur in many stories from classical mythology to science fiction.[2] Characters take various attitudes towards people, whilst not realizing that they are themselves the very people concerned.

Stories like these do not just appear in fiction; they appear in philosophy too. Let us start with two examples (we shall see more later). John Perry recalls following a supermarket shopper who was spilling sugar from his shopping cart, only to realize that he, like Winnie-the-Pooh, was following his own trail. More fancifully, David Kaplan imagines seeing the reflection in a window of someone whose pants are on fire, whilst failing to realize that it is his own pants that are on fire.[3]

We can see that such stories might be engaging; but why are they *philosophically* interesting? The reason is that they highlight a difference between knowledge about the world, impersonally conceived, and knowledge about our place in the world. As David Lewis graphically puts it, they highlight the difference between the information given by a standard map, and that given by a map that is erected in a public place with a "You are here" arrow. One tells you about the nature of the world; the other tells you, in addition, how you fit into that world. Following Lewis, call the former *de dicto* knowledge, and the latter *de se*.

The challenge posed by such a distinction is to come up with an account of the *de se*. Standard theories find it hard to do so. Standard theories treat our knowledge (and our beliefs, desires, etc.) as involving attitudes towards *propositions*. There is some disagreement as to quite what propositions are, but on none of the main approaches is there an obvious way to understand *de se* knowledge.

One approach takes propositions to be structured entities. For Russell, the structured proposition corresponding to Kaplan's belief that a man's pants are on fire is the ordered pair of the man in question, and the property of being on fire. Crucially, given that the man in question is Kaplan, nothing changes when we turn to the proposition corresponding to Kaplan's belief that his own pants are on fire. The object of that belief is again the ordered pair of the man in question – that is, Kaplan – and the property of having burning pants.

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

There is more opportunity to make a distinction on Frege's account of structured propositions, which involves not just the objects and properties concerned, but also the ways that these things are thought of: in Frege's term, "modes of presentation." Here, then, the proposition corresponding to Kaplan's belief that that man's pants are on fire might be different from that corresponding to his belief that his own pants are on fire, since they might involve different modes of presentation of the same object. But what is the mode of presentation that one has of oneself? It doesn't look as though it could correspond to any normal description, since the same worries would apply: one could think that someone met that description, without realizing that one met it oneself. Frege concluded that "Every one is presented to himself in a particular and primitive way, in which he is presented to no one else."[4] But it is unclear quite what this primitive way is, and, as Frege acknowledged, it makes communication indirect: I cannot communicate my *de se* thought to you, since you cannot grasp it.

If Frege's approach leads to obscurity, the alternative approach that takes propositions to be unstructured is in even worse trouble. The central proposal here is Lewis's own, according to which propositions are classes of possible worlds. But if names and demonstratives are rigid, referring to the same thing in each world, then, if it is Kaplan's pants that are on fire, the class of worlds in which "that man's" pants are on fire is just the same as that in which Kaplan's pants are on fire. We have nothing to distinguish the *de se* belief from the *de dicto*.

In response to these difficulties Lewis proposes an account that, characteristically, is at once simple and radical.[5] Do not think of *de se* thoughts as attitudes to propositions at all. Think of them rather as self-ascriptions of properties. When Perry realizes that he is following his own trail, he self-ascribes the property of spilling the sugar; when Kaplan realizes that it is his own reflection that he can see, he self-ascribes the property of having flaming pants. Moreover, the *de dicto* can then be thought of as a special case of the *de se*: the case in which what is self-ascribed is the property of membership of a world.

Lewis's account thus has two parts. The first part, which is the focus of most of his discussion, and of most of the discussion that has followed, involves treating the objects of the attitudes, not as propositions but as properties. The second part, which is much less to the fore in Lewis's presentation and in the subsequent discussion, involves treating our attitude to these properties as that of self-ascription. I hyphenate the term to stress that this is a primitive notion for Lewis. To self-ascribe a property is not just to ascribe a property to oneself, as one might ascribe it to someone else. That would just give us back the question of what it is to think of oneself in the right way. After all, when Kaplan ascribes the property of having burning pants to the figure in the window, he does ascribe that property to someone who is indeed Kaplan; in one sense then, he ascribes the property to himself. What Lewis wants to insist is that he does not self-ascribe that property. The self and the ascribing do not detach.

I will argue that, as a result of underplaying its second part, the radical nature of Lewis's account has been largely missed. In particular, much recent literature has tried to incorporate his account simply by treating the objects of the attitudes as centered worlds, where a centered world is an ordered pair of a possible world together with a spatiotemporal location. But centered propositions just correspond to properties. So whilst they give us the right objects for the Lewis account, they go no way to providing the right attitude. (Moreover, I shall argue that they provide an unnecessarily baroque way of specifying contents.)

As a result of this neglect, many of the difficulties that Lewis's approach faces have been overlooked. The idea of primitive self-ascription is an obscure one. Our natural grasp on it is via the general idea of ascribing a property to an object, but this is exactly what we are not allowed to do here. Instead we have to treat the relation as fundamental, something that becomes increasingly hard to do when we consider first-person plural ascriptions, and ascriptions where the first-person pronoun is not in subject position. These difficulties do not render Lewis's account totally unworkable; but they do make it much more problematic than has been generally recognized.

25.1 Lewis's Method

Let us start by getting clear on exactly what Lewis takes himself to be doing, and on the techniques he uses. His main contention is that propositions cannot be the objects of the attitudes, since there are situations in which an agent has a great deal of propositional knowledge, and yet lacks some further *de se* knowledge. The idea then is that this further knowledge cannot be knowledge of a proposition, that is, it cannot be *de dicto*. (I will follow Lewis and treat talk of propositional belief and of *de dicto* belief as equivalent for now. The obvious problem with doing so is that many have thought that there exists a further category, *de re* belief, which is propositional without being *de dicto*. As we shall see, Lewis denies that this is really a separate kind of belief; but the issues there are tangential.)

One way of arguing for this makes use of examples like those that we have seen from Perry and Kaplan. These examples certainly establish that you can have some propositional knowledge about a situation and yet lack further *de se* knowledge about it. And the Kaplan cases show, beyond that, that you can have direct perceptual knowledge that a property obtains of a person who happens to be you, without knowing that you – you *yourself* as Castañeda would say – have that property. What the cases don't show, at least not without a number of further assumptions, is that this *de se* knowledge cannot be understood as propositional knowledge. For it could be that there is some way of deducing the *de se* knowledge from additional *de dicto* knowledge, if only one had enough of it. And if the *de se* knowledge is deducible from the *de dicto* knowledge, then presumably it in turn is *de dicto*.

This is where Lewis's approach is innovative. He wants to argue that you can have all of the *de dicto* knowledge that there is, and still lack *de se* knowledge.[6] For this he needs some new examples, since the Kaplan and Perry cases lack this feature. If you filled in the propositional knowledge of the subjects involved – gave them information about the angle of the reflection, say, or about the complete extent of the sugar trail – they would be able to deduce that it was their pants on fire, or their trail of sugar. That is not so say that they would have deduced such conclusions from entirely *de dicto* beliefs; that would still need to be investigated. But it does mean that, without further elaboration, the examples will not serve to show that one can have complete *de dicto* knowledge and lack *de se* knowledge.

The main example that Lewis gives to make this point is that of the two gods:

> Consider the case of the two gods. They inhabit a certain possible world and they know exactly which world it is. Therefore they know every proposition that is true in their world. Insofar as knowledge is a propositional attitude, they are omniscient. Still I can imagine them to suffer ignorance: neither one knows which of the two he is. They are not exactly alike. One lives on top of the tallest mountain and throws down manna; the other lives on the top of the coldest mountain and throws down thunderbolts. Neither one knows whether he lives on the tallest mountain or on the coldest mountain; nor whether he throws down manna or thunderbolts.[7]

This example is somewhat problematic, since it is rather hard to fill in the details. If the gods are able to choose their actions and then implement them, can they not infer who they are from seeing the results?[8] So let me give another of Lewis's examples in which the difficulties, if still present, are rather less to the fore. This example, originating from Perry, is that of Lingens, lost in the Stanford Library, who has such complete amnesia that he does not know who, or where, he is. As Lewis develops the story, Lingens has an alter ego – Lauben, let us suppose – who is lost, in a similar condition, in the Widener Library at Harvard. Let us further suppose that the Stanford and Widener libraries are even more comprehensive than they are now – their works include all *de dicto* knowledge – and that Lingens and Lauben have read and remembered every word. Let us also suppose that the libraries are qualitatively identical inside; not only are there no "You are here" maps on the walls, there are

no stamps in the books saying "Property of Widener" or "Property of Stanford," no Windsor chairs bearing crests emblazoned "Veritas" or "Die Luft der Freiheit weht" and so on. Then despite their complete *de dicto* knowledge, including their knowledge of there being two amnesiacs lost in different libraries, neither Lingens nor Lauben will know which they are.

Do they lack all *de se* knowledge? Surely not. If Lingens sneezes he knows that he sneezes; likewise for Lauben. (If the libraries really contain all *de dicto* knowledge, including reports of all the sneezing that goes on, then if one sneezes the other had better sneeze too; or else Lingens and Lauben would be able to work out who was who.[9]) And they will each know that they are either Lauben or Lingens. What each lacks is any *de se* knowledge that distinguishes him from the other, and so allows him to integrate his *de se* knowledge with the *de dicto* knowledge that is given to him by the books.

This is a powerful argument to show that *de se* isn't *de dicto*. It is reminiscent of Frank Jackson's famous *knowledge argument* for the non-reducibility of phenomenal knowledge to physical knowledge: Mary knows all of the physical facts of color science, but because she has never seen a red thing, she does not know what red looks like.[10] Of course in that case Lewis doesn't want to acknowledge the existence of objects of knowledge that are not physical; he wants to understand the apparent phenomenal knowledge as knowledge-how.[11] But in the case of the *de se* he does want to move from what intuitively look like different knowledge states to the conclusion that there must be different objects of knowledge.

Before we examine the details of that account, let me remark on a potential distorting feature of Lewis's argument. Reflecting on cases like these leads naturally to the thought that if the *de se* and the *de dicto* are two independent bodies of knowledge, what is often needed is what I'll term *breakthrough knowledge*: a piece of knowledge that enables the two to be connected. And, given the nature of the cases, it can seem that the breakthrough knowledge must be *de se* knowledge of spatiotemporal location. One of the gods realizes that he is the god on the highest mountain; Lingens realizes that he is in Stanford Library. Once they have this piece of knowledge, everything else follows: once the wall between the two bodies of knowledge is broken, each can flow into the other.

However, this focus on spatiotemporal location as distinctive of breakthrough knowledge is misleading. For a start it is not necessary. Any uniquely identifying piece of *de se* information would do the job. Lingens could discover that he is called "Lingens"; one of the gods might discover that he throws down thunderbolts. Second, for some creatures – permeable creatures who can share locations with others – it is not sufficient. More broadly, the idea that a single piece of breakthrough knowledge is enough to unify the *de se* and the *de dicto* is only true in the highly artificial cases that Lewis presents: cases in which the agents have complete and unified *de dicto* knowledge, complete and unified *de se* knowledge, and the ability to make the necessary inferences. Remove these conditions and spatiotemporal location will no longer always do the job. Perry might have had a GPS device that told him exactly where (and when) he was; that still wouldn't have allowed him to identify himself as the person making the mess, since he lacked the relevant *de dicto* knowledge that it was the person located where he was who was making the mess. In creatures who lack some *de dicto* knowledge then, coming to know one's spatiotemporal location provides no guarantee of breakthrough knowledge. This will be important when we examine centered worlds.

25.2 Lewis's Account of the *De Se*

Let us examine Lewis's account a little more closely. As we have seen, it involves two features:

1. the content of the attitude is a property;
2. the attitude taken towards the content is that of self-ascription.

As I have already stressed, taking self-ascription as primitive is crucial to Lewis's account. We normally think of ascription as a two-place relation: one ascribes a property to a thing. Self-ascription would then be the special case where the thing is the self. But that won't do the work here. If the self is just thought of extensionally, then we would have no way to distinguish the belief that one's pants were on fire from the belief that the pants of someone, who is you though you don't realize it, are on fire. So we would have, instead, to think that the self came with a certain mode of presentation. If we were to do that, then Lewis's account would collapse into a rather baroque version of Frege's.

So we have to think of self-ascription as a one-place relation: one simply self-ascribes a property. There is a real question whether we can make sense of this, but I'll postpone that till later. For now note how little discussion there is of the idea, either in Lewis's own work or in the subsequent literature.

It is perhaps not surprising that Lewis gave little fanfare to this second feature, for when he presented his account this was not the innovation. Implicitly at least, Lewis's account of our attitude to propositions had been much the same – to believe a proposition is to self-ascribe membership of a set of worlds[12] – so he understandably focused on what was new. Indeed, as mentioned above, Lewis now explicitly treats the *de dicto* as a special case of the *de se*: a *de se* belief in which the property that is self-ascribed is membership of a certain set of worlds.[13]

There is a second reason why the crucial role of self-ascription has been largely overlooked. In much recent work, Lewis's account has been understood in terms of *centered worlds*, where a centered world is a possible world together with a center, typically (though not in all accounts) a set of spatio-temporal coordinates.[14] On this way of thinking, the objects of the attitudes are classes of centered worlds. There seems to be no appeal here to a primitive relation of self-ascription. Indeed, there seems to be no appeal to properties as the objects of the attitudes either. So how do we get from Lewis's account to this? Let us examine the proposal in some more detail.

25.3 Centered Worlds

This explanation of the account in terms of centered worlds originates with Lewis himself, in his development of some comments from Quine. As Lewis puts it, the idea is that whenever we have a property we have a centered world: the centered world in which the object at the center has that property. Conversely, whenever we have a centered world we have a property: the property that is had by the object at the center. So instead of phrasing the account in terms of properties, we can phrase it in terms of centered worlds.[15]

The first thing to note is that, as Lewis presents things, the role for centered worlds is to stand in for properties. So we still need the idea that they are self-ascribed. That is something that is obscured in much of the subsequent literature, where centered worlds are often taken to play the role of propositions, that is, as things that are straightforward objects of belief, rather than of self-ascription. Introducing centered worlds does not change the fundamentals of the account; it just changes the way we describe it.

Why would we want to couch the description in terms of centered worlds rather than properties? A common idea is to view the center as akin to the "you are here" flag on Lewis's map. Or, less metaphorically, as David Chalmers puts it, "We can think of the centre of the world as representing the perspective of the speaker within the world."[16] Still, if this is going to help, we must get clear on what such a perspective is.

We need to put one possible misunderstanding immediately out of the way. The center cannot be the perspective that the thinker actually has on the world, the place where they are actually located. Of course in working out the denotation of indexical expressions we often need to know that: in

working out what a speaker denotes with a use of the term "Here" we need to know where they actually are. But this is not what we are doing when we use a centered world to represent the content of a belief state. After all, a thinker can make a mistake about where they are, in which case the center of the possible world will not be their actual location.

So would it be better to think in terms of where the speaker believes her perspective is? That is better, but it is still not quite right. To see this, consider some cases.

(1) Lingens discovers that he is in Stanford Library

This is easy to represent on the centered world approach. Since he knows all the *de dicto* facts, he will have narrowed the possible world that he is in down to one. When he discovers that he is in Stanford Library this will give him his spatiotemporal location, which can serve as the centre of his centered world. Self-ascribing the property of being at that location will enable him to make all of the other self-ascriptions, he will know his name, date of birth, and so on. The location thus serves as his breakthrough belief.

(2) Lingens is unsure whether he is in Stanford or Harvard.

Now we return to his state before the discovery. Here the case is like the previous one, except that there are two places in which Lingens thinks that he might be. So his belief is modeled by two centered worlds, one centered on Harvard, and one centered on Stanford. Each of the two provides a potential breakthrough point, in the sense that were he to self-ascribe that location, then that would serve to integrate his *de se* and his *de dicto* knowledge.

(3) The Consul knows only that he's in pain

We leave the upright Lingens to focus on the dissolute Consul. After hard nights out, the Consul sometimes awakens with no knowledge of who, or where, he is. For the short time before he opens his eyes, all he knows is that he is in pain. On the centered world approach, the content of his thought amounts to the full set of possible locations; it is just that each location is paired with a world in which there is a person in pain at that location. This brings out how odd the centered world approach is. We started with the idea that the Consul simply self-ascribes a single property; he believes far less than Lingens. But that is cashed out using a far larger content: not just a huge array of different worlds, but also a huge array of different centers. Is this a problem? I think not, but it does make clear what these centered worlds are doing in Lewis's proposal. They are not meant to be psychologically realistic, in the sense that agents would actually have an idea of each of the worlds and the centers. Rather, they simply work to describe, or index, the various thoughts that the agent might have.[17]

(4) Felicité thinks she may be the Holy Ghost

Felicité has come to think that she may be the Holy Ghost. That is, although she thinks she might just be Felicité, an ordinary human being, she thinks that it is possible that she is actually the Holy Ghost, currently seeing the world from Felicité's perspective. She thinks that Felicité has a simple location but the Holy Ghost does not. Perhaps he is everywhere or perhaps he is outside time and space altogether. If he is everywhere, she thinks that it is possible that he is the only being that is everywhere (assuming that he is not different from the Father and the Son, something that Felicité finds it hard to get clear on); but being convinced of the reality of the beings of the Dark Side, she is

far from confident of that. How then do we characterize her belief using centered worlds? The possibility that she is Felicité is easily handled: that is just the class of centered worlds whose center is wherever she takes Felicité (and only Felicité) to be (note that on her view, only Felicité is at just that one place). But the possibility that she is the Holy Ghost cannot be captured using a spatiotemporal centre, for she does not believe that there is any centre that uniquely locates the Holy Ghost.

In response to worries of this kind, Lewis suggests that we should take centers to be, not spatiotemporal locations, but, rather, individuals. Will that always do the job? Shen-Yi Liao has pointed out that, with beliefs involving time travel, it would not.[18] He looks at various other possibilities for the role of centre that might do better. However, it seems to me that to try for something that is guaranteed to work is to raise the standard impossibly high. It is to seek a *characteristica universalis*: a universal conceptual framework that has the resources to express, or at least to accommodate, any possible thought. It is far from clear that we will be able to do that, or that we need to. All that we need is to provide, for each centered world, a center that will serve as a breakthrough point, in the following sense: if the thinker self-ascribed the property given by that center (whether location or some other property), then that would enable the thinker to self-ascribe the other properties that are true of that world.

If we generalize the notion of center in this way, we might think that whenever a thinker gains *de se* knowledge, that is because they self-ascribe location of some such centre. But even that is not true. It is an impression generated, once again, by an over-restricted focus. To see this consider again one of the examples with which we started.

(5) Perry comes to realize that he is spilling the sugar

Assume, once more, that Perry knows exactly where and who he is. What then does he come to know when he realizes that he is spilling the sugar? It is not that he distinguishes between centers, accepting some and ruling others out. Rather, on the natural understanding at least, he distinguishes between worlds. That is, amongst the class of centered worlds with the same center, he comes to reject those in which the person located at that center – that is, he, Perry – lacks the property of spilling the sugar.

So, in working with the centered world approach, we are left with a very complicated picture. When we try to identify what is distinctive about *de se* knowledge we cannot in general do so by pointing to the role of the centers; indeed, it is far from clear quite how we would mark out the class of the *de se*. In contrast, when working with Lewis's original formulation, the *de se* is clearly marked out: a *de se* belief is one that involves the ascription of a property that is not membership of a world. My suspicion is that the enthusiasm for the centered world formulation stems from a preoccupation with truly self-locating beliefs: the kind that feature in the Two Gods case. But, as we have seen, not all *de se* beliefs are like that. We would do much better to stick with Lewis's original formulation.

25.4 Attitudes *De Re*

Traditionally philosophers have recognized at least three classes of belief. There is the pure *de dicto*, which we can think of as composed of knowledge – or, at least, belief – by description. There is the *de se*. And then there is the *de re*: belief that involves direct acquaintance with an object, but which need not be *de se*. When Kaplan sees, unknowingly, that his pants are on fire, he has direct (or at least, merely reflection-mediated) *de re* perceptual knowledge of the flaming pants; but he lacks the *de se* knowledge that they are his.

Lewis's approach is designed to capture the *de se*. The *de dicto* comes as the special case in which I self-ascribe membership of the relevant worlds. But what of the *de re*? A *de re* belief is a belief about some particular thing, for instance something that I can identify by seeing it or touching it. In that sense it goes beyond the *de dicto*, and so cannot be reduced to the *de se* by the device of merely having a world in common. But on the other hand, many *de re* beliefs are not directly *de se* beliefs. If I say, by looking at you, that you are tall, it looks as though I have made an ascription of a property to you. It doesn't look, on the face of it, as though self-ascription comes into it at all.

Here Lewis's approach is to divide and conquer. *De re* beliefs are not pure mental states he says. Indeed, on Lewis's view, nothing that involves a relation to an external object can be a belief, since beliefs, like all mental states, are in the head. Instead *de re* beliefs are an amalgam of a mental state and a relation to an object. The mental state is *de se*.

Let us here grant to Lewis this composite characterization of the *de re*; whilst much could be said against it, and against the idea that belief must be in the head, my focus will be on the mental state that Lewis thinks makes up part of the *de re*. How can this be understood as *de se*?

Lewis's idea is that to believe *de re* of an object Y that it has a property X is "to ascribe X to Y under some suitable description of Y." Most of Lewis's discussion is concerned with elucidating what a "suitable description" is (it is either something that captures the essence of Y, or is a relation of acquaintance to Y). But prior to this comes Lewis's account of what it is for a subject to ascribe a property X to an individual Y under description Z. This obtains iff:

1. the subject bears the relation Z uniquely to Y; and
2. the subject self-ascribes the property of bearing relation Z uniquely to something which has property X.

This makes clear the composite nature of the *de re*. The first condition is not mental; the second is: it is a *de se* mental ascription. To make this more concrete: whilst the character in Kaplan's example does not self-ascribe the property of wearing the burning pants, he does self-ascribe the property of seeing them.

If this account is to work, every case of *de re* belief must be a case in which the subject self-ascribes a particular unique relation to the thing in question, that is, to the *res*. But why should we think that that is so? Couldn't a subject have a belief about the *res* without having any belief about the relation they bear to it? Or, equally worrying for Lewis's account, does the subject have to believe that the relation is unique: couldn't a subject have, and take themselves to have, two different *de re* beliefs whilst self-ascribing no difference in the relation that they bear to the two different things concerned? At the very least, to deny this is a substantial assumption. But I shall not pursue that worry here. Let me instead turn to examine some of the other potential costs of Lewis's approach, those that stem from the idea of primitive self-ascription.

25.5 Some Costs of Primitive Self-Ascription

I doubt that there is anything like a knockdown argument against the very idea of primitive self-ascription. But it is certainly odd; at least as odd as anything that Frege proposed. Indeed, like Frege's account it makes the *de se* essentially incommunicable.[19] My own take on it tends to run back, in a way that is hard to resist, to the idea that I am simply ascribing properties to one particular thing in the world, namely myself. That is why it makes sense to ascribe to myself the same sorts of properties that it makes sense to ascribe to other people; and why – a point we shall come to shortly – it makes

sense to make first-person plural ascriptions, where the same property is simultaneously ascribed to myself and to others.

In contrast Lewis has to think that self-ascription is not only different, but fundamental. Since everything is at base a self-ascription, that is where we must start. This takes him to a position that is almost Cartesian in flavor: we start with self-ascription, and then build our way out. Indeed, self-ascription may be the best way to understand Descartes' position. It is sometimes complained that Descartes is wrong to start the cogito with the premise "I think"; that suggests an independent take on the self as a self-standing object, which mere thought cannot give. All he is entitled to, the complaint continues, is the weaker premise "There is thinking."[20] But that seems too weak. Suppose the subject both thought and worried: then it should follow not just that there would be both thinking and worrying separately going on, but that they would be somehow coinstantiated.[21] But how can we make sense of that if we are not allowed to posit a self-standing subject? Here the idea of self-ascription might help us. Both thinking and worrying are self-ascribed, and so their conjunction can be self-ascribed, without the need to posit a subject.

Whilst such a starting point will be unattractive to those who reject Descartes' approach,[22] it does seem that we can get our minds around it. But when we look to more complex examples, that is less clear. To see this, let us start with the first person plural. If I say, of you and me:

We are hungry.

it seems that I have ascribed the same property – hunger – to each of us, and that the ascriptions are of the same form. Equally, if I can hear that one of our stomachs is rumbling but know not which, I can wonder whether to ascribe the property to you or to me.

On Lewis's account, none of this is straightforward. When I say that we are hungry, it might intuitively seem that I make two ascriptions, one to me and one to you. Translated into Lewis's account, these become two very different self-ascriptions. One of them is of the property of being hungry. The other (assuming this is a *de re* ascription) is of the property of inhabiting a world in which the person I'm acquainted with in a certain way is hungry. And likewise, when I wonder whose stomach it is that is rumbling, I wonder, not to whom to ascribe the property of having a rumbling stomach, but which of two very different properties to self-ascribe. As I say, to point this out is not to provide anything like a knockdown argument, and those sympathetic to the balance book method of doing philosophy might well judge that the economy provided by having just one sort of belief overcomes the inelegance of apparently similar contents coming out differently. But I am wont to use Lewis's response to Quine against himself: "I protest that the advantages of uniform objects are not to be lightly forsaken."[23]

Is there then another way that we might understand first person plural ascriptions? An obvious approach is to think that the "we" picks out a group, in which the speaker is included. Then rather than a conjunction of self-ascriptions, we could have one: self-ascription of membership of a group with a certain property. We can see how that might work for truly collective cases in which the group really does possess a property that is not possessed by the members individually, such as:

We carried the wardrobe upstairs.

It could also perhaps handle truly distributive cases, like our original "We are hungry." Here the idea is that one self-ascribes membership of a group each of whom has that property. What is less clear is how one would accommodate the so-called cumulative readings, like:

We ate seven pizzas between us.

This isn't collective; unlike wardrobe carrying, pizza eating, is typically something that is done individually, even if in the company of others. Equally though it is not distributive; it is not true that each member of the group ate seven pizzas.[24] So how are we to understand it as a single self-ascription to a group?

Such cases are rather hard to accommodate on any account though, so they should not count too heavily against Lewis. But there are other cases in which the plural pronouns really do become problematic for Lewis when they are quite straightforward for more traditional accounts. These occur when we combine them with another kind of difficulty: cases in which the first person pronoun is not in subject position. In general, to understand such sentences as self-ascriptions will require radically reconceptualizing the sentence in question. Sometimes this is not too complicated. A sentence like

Hilary chose me.

will involve self-ascribing the property of being chosen by Hilary, and

Hilary chose Rose to assess me.

will involve self-ascribing the property of being the person whom Hilary chose Rose to assess. But what of

I chose myself.

which seems to involve self-ascribing the property of self-choosing? Or, worse still, what of

I chose myself to assess myself.

a doubtless corrupt but nonetheless comprehensible action, which seems to involve self-ascribing the property of self-self-assessing-choosing?

When we combine these cases with singular and plural pronouns things become even more intractable:

I chose us to assess me.

which I cannot manage to parse in the form of a self-ascription. If possessive pronouns also involve self-ascription, the same difficulties will arise in much more mundane cases:

I cleaned myself with my handkerchief.
I defended our house with my friends.

Perhaps there is a way to parse these, but the underlying worry should be clear. To understand these sentences in terms of self-ascription requires not simply rearranging the ideas, but coming up with novel properties, properties of self-choosing and so on. If self-ascription really is primitive, then it seems that it is with such properties that we start. That is not an incoherent idea; but it is very hard to believe.

I have confined my discussion here to knowledge and belief. But Lewis wanted to extend it to wanting. Indeed, since pretty much any propositional attitude can come with a *de se* content – not only can I believe myself to be a certain way or want myself to be a certain way, I can also imagine

myself, intend myself, fear myself, and so on, to be a certain way – it seems that we will have to treat all of the propositional attitudes in the same way. That means that the problems I have discussed here will ramify. Not only will we need a primitive notion of self-ascribing, we will also need a primitive notion of self-wanting (i.e. wanting oneself to be a certain way), self-imagining (i.e. imagining oneself to be a certain way), self-intending (i.e. intending oneself to be a certain way), and so on; and each of these will in turn enter into the same problems with plural subjects, non-subject position, and so on.

25.6 Conclusion

Lewis's account of the *de se* is both more radical, and more unintuitive, than has been appreciated. Lewis, of course, was always prepared to accept the unintuitive if it made for elegant theory. What is less clear is how elegant the theory can be made to be.

Acknowledgments

Thanks to audiences at MIT, Paris, and Cambridge, and to Rae Langton, Steve Yablo, and Seth Yalcin.

Notes

1 Milne 1926, Ch. 3.
2 In one version of the story of Jason and the Argonauts, Jason suggests that a would-be assassin should be sent to fetch the Golden Fleece without realizing that he is the would-be assassin. In a rather more recent story, Robert Heinlein (1959) has a time-traveling character consistently failing to recognize himself.
3 Kaplan 1989.
4 Frege 1984.
5 Lewis 1979. See also Lewis 1986 and 1983b.
6 In discussing an example of Castañeda's, Lewis writes: "To support this claim [that attribution of knowledge *de se* is not equivalent to any attribution of knowledge *de dicto*] we need only find a case in which the editor knows well enough which of the worlds is his without knowing whether he is among the millionaires of his world" (1979, 139–40).
7 Lewis 1979, 139.
8 For a good attempt to make sense of the possibility, see Stalnaker 1999, 56, n.6.
9 Don't push this thought too far, or we end up with similar problems to that of the two gods.
10 Jackson 1982.
11 In not knowing what red looks like, Mary lacks the ability to identify the color of things just on the basis of their appearance. See Lewis 1988.
12 See Lewis 1979, 137.
13 As a result it is unclear that Lewis can give a plausible account of beings that can have only *de dicto* beliefs: that lack the concept of the self. Yet mightn't there be some? (Perhaps some of those creatures that fail the mirror test.)
14 For a summary of the work using centered worlds, and of the roles to which they have been put, see Liao 2012.
15 Lewis 1979, 147.
16 Chalmers 2004, 60.
17 For Lewis's description of this as his aim, see the Postscript C to "Attitudes" (1983a, 158–9). Of course we might still want something that is more psychologically realistic, which would give us reason for avoiding the centered world approach.

18 Liao 2012.
19 See Stalnaker 1999 for some discussion of this. Note though that I am not suggesting that we should abandon Lewis's account in favor of Frege's. Each of them makes sense only in the context of a very different account of the nature of the content of thought.
20 This worry is normally traced to Georg Lichtenberg, *Waste Book* K76 (c.1794); translation in Lichtenberg 2012.
21 This coinstantiation concern is raised by Bernard Williams (1978, 95–100).
22 The classic response here is Quine 1969.
23 Lewis 1979, 147.
24 See Landman 1995. Thanks to Friederike Moltmann and Irene Heim for discussion here.

References

Chalmers, D. (2004). The foundations of two-dimensional semantics. In M. Garcia-Carpintero & J. Macia (Eds.), *Two-dimensional semantics*, pp. 55–140. Oxford: Oxford University Press.

Frege, G. (1984). The thought. In B. McGuinness (Ed.), *Collected papers on mathematics, logic and philosophy*, pp. 351–72. Oxford: Blackwell.

Heinlein, R. (1959). By his bootstraps. In R. Heinlein, *The menace from earth*. New York: Gnome Press.

Jackson, F. (1982). Epiphenomenal qualia. *Philosophical Quarterly*, 32, 127–36.

Kaplan, D. (1989). Demonstratives. In J. Almog, J. Perry & H. Wettstein (Eds.), *Themes from Kaplan*, pp. 481–564. New York: Oxford University Press.

Landman, F. (1995). Plurality. In S. Lappin (Ed.), *Handbook of contemporary semantics*, pp. 425–57. Oxford: Blackwell.

Lewis, D. (1979). Attitudes *de dicto* and *de se*. *Philosophical Review*, 88, 513–43. Reprinted with postscripts in Lewis (1983a), pp. 133–59. Page references are to the latter version.

Lewis, D. (1983a). *Philosophical papers*, Vol. I. New York: Oxford University Press.

Lewis, D. (1983b). Individuation by acquaintance and stipulation. *Philosophical Review*, 92, 3–32. Reprinted in Lewis (1999), pp. 373–402.

Lewis, D. (1986). *On the plurality of worlds*. Oxford: Blackwell.

Lewis, D. (1988). What experience teaches. *Proceedings of the Russellian Society*, 13, 29–57. Reprinted in Lewis (1999), pp. 262–90.

Lewis, D. (1999). *Papers in metaphysics and epistemology*. Cambridge: Cambridge University Press.

Liao, S. (2012). What are centered worlds? *Philosophical Quarterly*, 62, 294–316.

Lichtenberg, G. (2012). *Philosophical writings*. Ed. S. Tester. Albany: State University of New York Press.

Milne, A.A. (1926). *Winnie the Pooh*. London: Methuen.

Quine, W.V.O. (1969). Epistemology naturalized. In W.V.O. Quine, *Ontological relativity and other essays*, pp 69–90. New York: Columbia University Press.

Stalnaker, R. (1999). *Our knowledge of the internal world*. Oxford: Oxford University Press.

Williams, B. (1978). *Descartes*. Harmondsworth: Penguin.

26

Counterfactuals and Humean Reduction

ROBERT STALNAKER

26.1 Introduction

David Lewis started his career with two major constructive projects, each aiming to do something that his teachers had argued cannot be done. First, in his dissertation and the book *Convention* that grew out of it, he argued that, despite what W.V. Quine believed, one *can* make sense of truth by convention. Then, in his book *Counterfactuals*, he argued that, despite what Nelson Goodman had reluctantly concluded, one *can* give a reductive analysis of counterfactual conditionals. My topic is this second project: I will try to spell out the ways in which Lewis's aim was a reductive analysis, and the ways that it compares and contrasts with Goodman's project. It's not that Lewis argued that one could do exactly what Goodman gave up on doing; the framework in which he set the problem up was strikingly different. But both Goodman and Lewis were animated by Humean skepticism about natural necessity. What needed to be analyzed away was a family of concepts that apparently described or implied real relations and connections between distinct events. For both Goodman and Lewis, causal dependence and independence, capacities, dispositions, potentialities, and propensities were problematic, and so in need of analysis. But pinning any project of analysis down requires getting clear about the linguistic and conceptual resources that are available – about the unproblematic base to which the family of problematic concepts was to be reduced. A reductive project may fail, not just because no satisfactory analysis could be found, but also because the distinction between the problematic concepts and the unproblematic base was not sufficiently clear or well motivated. In this chapter I will argue that Lewis's reductive project does not succeed, partly for this kind of reason. If I am right, this raises the same question that the acknowledged failure of Goodman's project raised: what is the alternative to a reductive analysis? If we conclude that no reductive analysis is to be had, but still find a concept or family of concepts philosophically problematic – in need of some kind of explanation – what do we do?

Here is my plan: I will start, in section 26.2, with Goodman's project, saying what his aim was, spelling out the resources that he allowed himself to use in order to accomplish this aim, and explaining why the aim could not be accomplished with these resources. I will conclude this section by

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

considering Goodman's response to the acknowledged failure of his initial attempt at analysis: how he proposed to redefine his Humean project. In section 26.3, I will sketch Lewis's project, which has two parts: an abstract formal semantic analysis of conditionals, and an attempt to explain the primitive parameters of the semantic models in a way that is austere enough to satisfy the Humean strictures, but rich enough to provide an intuitively satisfactory account of counterfactual conditionals, and other concepts in the family that the Humean argues are in need of explanation. In the next two sections I will look more closely and critically at the conceptual resources that Lewis used to characterize the base to which counterfactuals, laws of nature, causation, dispositions, capacities, propensities, and potentialities were to be reduced: in section 26.4 at the thesis of Humean supervenience, and in section 26.5 at Lewis's notion of a natural property. I will argue that Lewis does not succeed in identifying a suitable base for a reduction of counterfactuals and other causal notions, and then conclude, in section 26.6, with a brief remark on the problems that remain if we reject the project of Humean reduction.

26.2 Goodman's Project[1]

One might think that if certain concepts are problematic, then there is some problem about them. What are the problems about dispositions, counterfactuals, and possibilities that, according to Goodman, "are among the most urgent and most pervasive that confront us today in the theory of knowledge and the philosophy of science"? In a famous passage in *Fact, Fiction and Forecast*, Goodman gave the following nonanswer to this question:

> A philosophical problem is a call to provide an adequate explanation in terms of an acceptable basis.... What intrigues us as a problem, and what will satisfy us as a solution, will depend upon the line we draw between what is already clear and what needs to be clarified.... In the absence of any convenient and reliable criterion of what is clear, the individual thinker can only search his philosophical conscience. As is the way with conscience, it is elusive, variable, and too easily silenced in the face of hardship or temptation. At best it yields only specific judgments.... Indeed this talk of conscience is simply a figurative way of disclaiming any idea of justifying these basic judgments. (1983, 31–2)

If the kind of explanation one gives is a reductive analysis, then one does not need to say what the problem is in order to solve it. Whatever is problematic about some concept, the problem will disappear if one gives an eliminative definition of the concept in terms of notions that do not have that problem, whatever it is. But if the project of reductive analysis fails (as it usually does) then one will need to identify the problem or problems in order to see what kind of explanation, short of reduction, might solve it or them. Identifying the problems might also help to make clear what will count as an acceptable basis for the philosophical explanation. The terms in which one explains some problematic concept must be clear, with respect to the particular problems at issue, but an analysis or explanation might still be successful, even if the notions one starts with are themselves problematic, but face different problems. (For example, one might think that the concepts of truth, belief, and justification all face serious philosophical problems, but still think that if the traditional justified true belief analysis of knowledge had not fallen to counterexamples, it would have provided a solution to problems about skepticism.)

It is clear enough why *Hume* thought that notions involving natural necessity were problematic and in need of analysis. All legitimate ideas are copies of, or at least analyzable in terms of, sense impressions, and there is no sense impression corresponding to the idea of causation. Goodman's worries about counterfactuals were not tied to Hume's specific doctrine about the empirical basis for

concepts, or to the verificationist doctrine that was its twentieth-century descendant, but they are worries that have their source in empiricism. The logical empiricist project of explaining theoretical scientific notions in terms of more directly observational notions was a dauntingly difficult one, but it was evident to those pursuing this project that it would be a lot easier if one were allowed to include a counterfactual conditional operator among the logical resources used for such analyses (one of the temptations that the puritanical Goodman quotation above warns us to resist). Strictly extensional logical resources were unproblematic. Intensional connectives and operators threatened to smuggle non-empirical content into the concepts analyzed in terms of them, but if one could given truth conditions for sentences involving such a connective, using only extensional logical resources, that would justify using counterfactuals in one's explanations of the relations between theory and observation. The general form of analysis that Goodman started with was this: a counterfactual conditional, ($\phi \mathrel{\square\!\!\rightarrow} \psi$), is true if and only if ψ can be deduced from ϕ, conjoined with a true set of laws of nature and with a suitable set of true sentences (the relevant conditions). Filling out this skeleton of an analysis required two things: first, an account of laws of nature, distinguishing them from mere accidental generalizations, and, second, a distinction between suitable and unsuitable factual truths – between those truths permitted as premises in the derivation of the consequent and those that are not. Goodman considered the two problems separately, but he might have put them together into the single problem of identifying the contingent truths that are suitable as additional premises in a derivation of the consequent from the antecedent. It was obvious from the start that one couldn't allow all truths to be premises without trivializing the analysis, since in the case of counterfactuals, the negation of the antecedent is true, and so a set of all truths plus the antecedent will be inconsistent, making every counterfactual true. Goodman tried on various restrictions: the relevant conditions must be compatible with the antecedent, compatible with both the consequent and the negation of the consequent, etc. It was also proposed to add a negative condition – that there *not* be a suitable set of premises which, when conjoined with the antecedent was sufficient to derive the negation of the consequent. Each attempt could be trivialized by showing that it either made all counterfactuals true, or none of them. Then, as a rhetorical move preliminary to giving up, Goodman proposed that the set of relevant conditions be required to be not only consistent with, but *cotenable* with the antecedent, where a sentence is cotenable with the antecedent if and only if it is not the case that it would be true if the antecedent were true. But, as he then pointed out, with this analysis, "we find ourselves involved in an infinite regress or a circle, for cotenability is defined in terms of counterfactuals, yet the meaning of counterfactuals is defined in terms of cotenability" (16). At this point, he gave up. Six years after the publication of his initial paper, he was still smarting from "the unsullied record of frustration" and the "years of beating our heads against the same wall and of chasing eagerly up the same blind alleys," and so was ready to try something different. Before looking at the new strategy that he proposed, let me make two general remarks about the failed project.

First, Goodman did not give an argument that his project *had* to fail. He just showed that none of the particular proposals that he tried were successful. But he could have given what is close to a proof that the resources he allowed himself were bound to be insufficient. In specifying relevant conditions (the class of sentences that were admissible as premises to be added to the antecedent in a derivation of the consequent) one is allowed to appeal only to (1) the truth-value of the sentences (only truths were allowed) and (2) the logical relations between the candidate sentence and the antecedent and consequent sentences. One could require that to be a relevant condition, the candidate sentence must be compatible or incompatible with or, entail or be entailed by the antecedent, or the consequent, or their negations. But it was not permissible to appeal to the *syntactic* structure of the sentences (for example, whether they were atomic, disjunctions, or negations), since that is an arbitrary matter of formulation. Using a modeling tool that Goodman would not have found congenial, but which is still suitable for picturing the information he allowed himself for his reductive project, we can represent

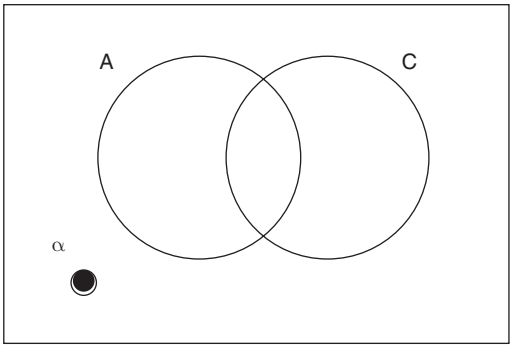

Figure 26.1

the situation with a Venn diagram with just two circles, one for the antecedent A and one for the consequent C, and a point α to represent the actual world (Figure 26.1).

All of the four regions distinguished in the diagram will be open, since we assume that the antecedent and consequent are logically independent. The point α is outside of both the A and C circles, since we assume that A and C are both in fact false. The diagram then contains all of the information that might be used to decide whether any arbitrary third circle is an admissible relevant condition, and so the diagram contains all of the information that might be used to determine whether the counterfactual, (A $\Box\!\!\rightarrow$ C) is true or false in the actual situation, α. But obviously, the diagram is the same for any two sentences A and C that are both false, and logically independent. Without adding some additional structure, there is nothing that could distinguish true from false counterfactuals.

The second point to note is that Goodman might have paused after offering his circular cotenability analysis, but, before turning away, to consider the consequences that this proposal has for the logical structure of counterfactuals. While failing to be a reductive analysis, the proposal is not thereby empty. Suppose one were to take the counterfactual to be primitive, and interpret the two parts of the circular analysis (cotenability defined in terms of the counterfactual, and the counterfactual defined in terms of cotenability) as postulates. There are a few details to pin down to make this suggestion precise, but, on a straightforward way of spelling it out, the result would be a structure that is essentially the same as that of the kind of abstract semantics that was first proposed more than 20 years after Goodman floated the cotenability proposal. Fixation on the general philosophical strategy of reductive analysis led Goodman, and everyone else at the time, to ignore any consideration of a line of inquiry that later had interesting results, and helped to sharpen Goodman's problem, if not to solve it.

So the exploration of the logical structure of the counterfactual was not part of Goodman's new direction. What was the strategy that he proposed to adopt after the failure of the initial project? His suggestion was that we give up on counterfactuals and start instead with dispositions.

> In dealing with counterfactuals . . . we are expressly concerning ourselves with a *form* of a statement, and the pattern of analysis we see is largely dictated by the structure of the conditional. This structure, although it promised at the outset to be a valuable aid, may actually have become a hindrance. The very disanalysis effected by returning to consider dispositional statements, which are indicative and simple in form, may free us to explore a better scheme of analysis. . . . I suspect that the problem of dispositions is really simpler than the problem of counterfactuals. (38–9)

Although it is clear that Goodman's new approach involves a switch in focus from counterfactuals to dispositions, it is less clear exactly what kind of analysis of them his new strategy requires. He suggests that we have already defused a part of the problem, simply by noting that "dispositional as well as manifest predicates are labels used to classify *actual* things" (59). They require no departure from standard extensional logic; they are ordinary predicates with ordinary extensions. He also notes the pervasiveness of dispositional predicates: adjectives like "hard" and "red" are as dispositional as words with "a tell-tale suffix like 'ible' or 'able' . . . Indeed, almost every predicate commonly thought of as describing a lasting objective characteristic of a thing is as much a dispositional predicate as any other" (41–2). Dispositions remain problematic since "they seem to apply to things in virtue of possible rather than actual occurrences" (42), but there are no merely possible occurrences. So the problem is "to explain how dispositional predicates can be assigned to things solely on the basis of actual occurrences" (42). He goes on to suggest that the move from a manifest predicate such as "flexes" to a dispositional predicate such as "flexible" involves a kind of conceptual projection. Things not under suitable pressure neither flex nor fail to flex, but we can extend the flex/fail-to-flex distinction to a wider range of cases. That is, the extension of the new predicate "flexible" coincides with the extension of "flexes" within the domain of things under suitable pressure, but also may be applied to things outside of that domain. We extend the distinction, Goodman suggests, by using the same principles we use to make *inductive* projections.

> The problem of dispositions looks suspiciously like one of the philosopher's oldest friends and enemies: the problem of induction. Indeed the two are but different aspects of the general problem of proceeding from a given set of cases to a wider set. The critical questions throughout are the same: when, how, why is such a transition or expansion legitimate? (58)

Goodman's attention then shifts to the problem of induction, and the task of constructing a substantive theory of confirmation. Counterfactuals and dispositions fade into the background and are not mentioned again.

Goodman started with an uncompromising line: certain notions are clear, others in need of explanation. The family of concepts that Hume found suspect – potentialities, capacities, and dispositions, causal dependence and independence, objective chance – are central to our science and epistemology, but this is reason to seek clarification, not to succumb to the temptation of taking them to be clear. They (according to his philosophical conscience) fall on the wrong side of a bright line. Furthermore, Goodman suggested that philosophical explanation requires analysis, by which he meant reductive definition. He was scornful of the idea of a partial definition: concepts are either fully defined, or allowed as primitives (see 46–7). But, by the end of the discussion of dispositions, there is at least some suggestion that the problem has been defused, not by analysis, but by explaining why it is okay to take some dispositional predicates as primitives. After all, if the simple paradigm cases of observational terms like "red" (not of a phenomenalist language, but of a common sense physical-thing language) are dispositional, then why should we think that dispositional predicates are necessarily problematic? We have to explain how we can have evidence that such predicates apply to things, even in circumstances where the disposition is not manifested, but that is a general epistemological problem. We have, Goodman argued, gotten rid of reference to merely possible things and occurrences, and perhaps that is enough to make at least some dispositions, with their limited counterfactual consequences, acceptable as they are. Compare Quine's discussion of counterfactuals and dispositions, which was heavily influenced by Goodman's ideas. Quine says there that "the subjunctive conditional has no place in an austere canonical notation for science" but that "we remain free to allow ourselves one by one any general terms we like, however subjunctive of dispositional their explanations (Quine 1960, 225).

26.3 Lewis's Project

As I noted, Goodman's circular cotenability analysis contained the seeds of an account of the abstract logical structure of counterfactual conditionals, but he did not stop to develop that account. Instead, he concluded that the focus on logical structure was a hindrance rather than a help. Lewis's project begins by returning the focus to the compositional semantic structure of conditionals. He constructs a formal language with primitive conditional connectives, and a semantics for interpreting them developed in the framework of possible worlds – a framework that did not exist when Goodman first attacked the problem of counterfactuals, and that he found extremely uncongenial after it was developed. Lewis's semantics gave truth conditions for the interdefinable "would" and "might" conditionals in terms of a parameter of a model – a three-place relation of comparative similarity between possible worlds (y is more similar to x than z is to x). The rough idea is that a conditional ($\phi \,\Box\!\!\rightarrow \psi$) is true in a world x if and only if ψ is true in all of the ϕ-worlds that are most similar to x.[2] The formal properties of any comparative similarity relation will help to explain the logical properties of the conditionals whose truth conditions are given in terms of it. For example, since we can assume (no matter what the relevant respects of similarity are) that there is no y more similar to x than x is to itself, the analysis will validate modus ponens.

Although it imposes and explains a logical structure, Lewis's abstract semantics does not, by itself, provide a Humean reduction, since truth conditions are given relative to a primitive parameter of the interpretation – the comparative similarity relation. The formal properties of a relation of this kind are not enough to identify it. Unless and until this relation is explained in a way that shows it to be unproblematic in the relevant sense, one has not solved the problem, but only given a framework that sharpens it by separating the logical problem of compositional structure from the substantive problem of specifying the respects of similarity that are relevant to the interpretation of counterfactuals, and so to other concepts that might be defined in terms of counterfactuals. So the substantive problem is where the action is, if a Humean reduction is to be successful, and what is required is not only a counterexample-free specification of the relevant respects of similarity between worlds, but also a clear and well-motivated identification of the resources that are admissible to the basis for the reduction. I think Lewis would agree with this assessment, and I will consider below his strategy for addressing the substantive problem.

In my own early paper on conditionals (Stalnaker 1968), which developed and defended a semantic analysis similar to Lewis's in the same possible-worlds framework, I distinguished the logical problem from the substantive problem (there, perhaps misleadingly, labeled "the pragmatic problem"[3]). While I thought there was more to be said about the substantive constraints on the selection function that encoded a comparative similarity relation, I did not think (then or later) that the problem was to find a reduction to concepts on some more basic level. And I argued that even without saying more about further constraints on comparative similarity, the abstract analysis does help to defuse one philosophical problem about counterfactuals – what I called "the epistemological problem" – a problem that may be part of the reason that the empiricist finds counterfactuals and related notions problematic. As I put the problem then:

> Counterfactuals seem to be . . . contingent, statements about unrealized possibilities. But contingent statements must be capable of confirmation by empirical evidence, and the investigator can gather evidence only in the actual world. How are conditionals that are both empirical and contrary-to-fact possible at all? (1968, 99)

This problem is similar to a worry expressed by Goodman about dispositions: "The peculiarity of dispositional predicates is that they seem to be applied to things in virtue of possible rather than actual occurrences" (1983, 42). This was my suggestion for defusing the problem: "There is no mystery to the fact that I can partially define a possible world in such a way that I am ignorant of some of the determinate truths in that world." This will happen if I define the world as a function of the actual world – in particular as similar to the actual world in some respect. For example, if I were to stipulate that the possible worlds I want to consider are those in which a certain yacht was two feet longer than it actually is, then it is easy to see how empirical evidence about the actual length of the yacht is relevant to what is true in this counterfactual world. If an abstract formal semantics of the kind that both I and Lewis proposed is right, then we can see that "conditionals do, implicitly and by convention, what is done explicitly and by stipulation" in an example such as the one about the yacht (Stalnaker 1968, 99). But even if this succeeds in defusing one general epistemological worry about counterfactuals and dispositions, there remains the problem of saying more, and of saying what more needs to be said, about the particular ways in which the selected counterfactual possibilities are similar to the world at which the counterfactual conditional is being evaluated.

In introducing his semantic analysis of counterfactuals, Lewis noted that the truth conditions he was going to propose were stated in terms of a primitive parameter of a model for the language, to be fixed within rough limits by the context of use, but he claimed that this parameter is a familiar one that had application that was independent of its use in the interpretation of counterfactuals. The claim that this parameter – the relation of comparative similarity – was familiar and independent of its use in the analysis of counterfactuals was supported by an example of an intuitive judgment of comparative similarity between cities ("Seattle resembles San Francisco more closely than it resembles Los Angeles"). He granted that possible worlds are usually bigger than cities, and may differ in a wider variety of ways, but

> still, any problems posed by my use of comparative similarity differ only in degree, not in kind, from problems about similarity that we would be stuck with no matter what we did about counterfactuals. Somehow we *do* have a familiar notion of comparative overall similarity, even of comparative similarity of big, complicated variegated things. (Lewis 1973, 92)

There are two problems with this reliance on the familiarity of the intuitive idea of overall similarity; one of them has been often noted, but the other is, I will suggest, the more serious problem, if the analysis is to be understood as part of a reductive account of notions that are problematic from a Humean point of view. The first problem is that if the notion of comparative similarity between worlds that is relevant to the interpretation of counterfactuals is the familiar one that we are presupposing when we judge that Lewis's claim about Seattle, San Francisco, and Los Angeles seems clearly true, then the analysis will face many counterexamples. Take any case where it seems intuitively that a certain choice not taken would have changed the course of history in dramatic ways if it had been taken. Possible worlds where the counterfactual choice is made, but the dramatic changes do not take place and things somehow manage to proceed in pretty much the way they actually did will seem, intuitively, to be more similar to the actual world. So if the truth of the counterfactuals depends on what happens in the most similar worlds (in this intuitive sense), the intuitively true counterfactuals will come out false (see Fine 1975). Lewis responded to this problem by saying that he never meant to suggest that the particular respects of similarity that are in play in the example of the cities are the ones that are relevant to counterfactuals. The point of the example of the cities, he said, was just to respond to a general skepticism about the notion of similarity (see Lewis 1986a, 52–5). In later work, Lewis made a detailed proposal about the respects of similarity that are relevant to the

interpretation of counterfactuals and he expressed surprise that he was interpreted as changing his account, rather than filling in more details that are relevant to one standard kind of context. But the discussion of intuitions about overall similarity in the book did seem to put more weight on that notion than it can bear.

The more serious problem with the reliance on some more or less intuitive notion of similarity is that it fails to reveal whether or not the respects of similarity that we are implicitly relying on include similarity with respect to features of the world that are suspect from a Humean point of view. One might think that an analysis in terms of similarity would be acceptable since the Humean is not skeptical about either purely spatial–temporal relations or relations of similarity and difference, which can be explained in terms of the sharing of properties. But the acceptability of a similarity relation depends on the acceptability of the properties that are the relevant respects of similarity. One would of course not have a Humean reduction if the properties shared by things that are similar in the relevant respect were powers, dispositions, capacities, and potentialities. Lewis's more detailed specification of the relevant respects of similarity aims to address this problem, as well as to avoid counterexamples, but doing so requires not only spelling out particular criteria for comparative similarity, but also showing how to describe the things to be compared – those "big, complicated, variegated things," possible worlds – in a way that is untainted by the problematic notions. That requires a general metaphysical theory, but of course Lewis has one.

26.4 Humean Supervenience

Lewis characterizes the general thesis that motivates his reductionist program this way:

> Humean supervenience is named in honor of the great denier of necessary connections. It is the doctrine that all there is to the world is a vast mosaic of local matters of particular fact, just one little thing after another . . . We have a geometry: a system of external relations of spatio-temporal distance between points, maybe points of space-time itself, maybe point-sized bits of matter or aether or fields, maybe both. And at those points we have local qualities. And that is all. There is no difference without a difference in the arrangement of qualities. All else supervenes on that.

Lewis holds that Humean supervenience is a contingent thesis, and he takes his task not to be to establish its truth, but to rebut objections to it by philosophers who

> claim that one or another commonplace feature of the world cannot supervene on the arrangement of qualities. . . . Being a commonsensical fellow (except where unactualized possible worlds are concerned) I will seldom deny that the features in question exist. I grant their existence and do my best to show how they can, after all, supervene on the arrangement of qualities. (Lewis 1986a, xi)

The plan for doing this follows a careful order, to avoid circularity: first, Lewis offers an account of laws of nature, following an old suggestion of Frank Ramsey that the laws are those universal generalizations "that achieve an unexcelled combination of simplicity and strength." The "best system" analysis of laws needs further development, as Lewis would agree, but he argues that the criteria of simplicity and strength are "safely noncontingent," and so unproblematic from a Humean point of view, and that the rest is just regularity. I am skeptical about both the clarity and the noncontingency of the criteria of simplicity and strength, but this is not my main worry, so I will grant the Humean seal of approval that Lewis places on his account of laws. The next step, and our main concern here, is the specification of the respects of similarity that are relevant to the interpretation

of counterfactuals – a specification that makes appeal to the previously explained notion of law. From there he goes on to the star of Hume's family of problematic concepts, the concept of causation, which is analyzed in terms of causal dependence, which is analyzed in terms of counterfactuals.

Here are the proposed criteria of comparative similarity, given as an ordered list:

1. It is of the first importance to avoid big, widespread, diverse violations of law.
2. It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
3. It is of the third importance to avoid even small, localized, simple violations of law.
4. It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly. (Lewis 1979, 47–8)

Lewis grants that there remains "plenty of unresolved vagueness" in this list of priorities, but he claims that it says enough to avoid the counterexamples that depend on the mistaken assumption that the analysis is based on an intuitive notion of overall similarity. My concern here is not whether counterexamples are avoided, but whether the specification appeals only to features of a possible world that satisfy the Humean strictures. As I said, I will set aside worries about the "best system" account of laws. The question I want to focus on is whether the defender of Humean supervenience can give a satisfactory account of particular fact, and of perfect match of particular fact. These facts, according to the specification of Humean supervenience, are constituted by the instantiation of intrinsic qualities by point-sized momentary objects, or perhaps by spacetime points themselves. But what are these qualities? For Hume, they would be simple ideas that are copies of impressions, but Lewis's version of the thesis disclaims any commitment to the phenomenalist aspects of Hume's metaphysics. ("It is no part of the thesis that the local matters are mental"; Lewis 1986a, ix.) On Lewis's metaphysical picture, the local qualities exemplified in the actual world are the fundamental properties that physical theory tells us about; more generally, in an arbitrary possible world, they are the *perfectly natural* properties that are instantiated in that world. It seems, at least prima facie, that properties of this kind (mass, charge, spin, the colors and flavors of quarks), unlike the phenomenal properties that Hume might have taken as fundamental, are dispositional properties whose causal powers are essential to them. The defender of Humean supervenience must claim, as Lewis does, that this prima facie appearance is mistaken. The world can be described in terms of fundamental properties that are metaphysically independent of the laws governing the things that exemplify them, and of the way that things with those properties behave. I will consider in the next section Lewis's general account of natural properties, and his argument that they are independent of their causal powers, but first let me conclude with a general comment about what is required for a satisfactory defense of a reductionist thesis.

There is a temptation to describe any dispute between a reductionist and an anti-reductionist (for example, between a materialist who argues that mental properties are reducible to physical properties and a dualist who denies this) in this way: the question is whether a description of all the facts of a certain kind (for example the physical facts) would be a *complete* description or whether there are features of the world over and above those facts. The reductionist who is not an eliminativist about a range of facts that appear not to be of the favored kind (the commonsensical fellow who doesn't deny the existence of the facts) is obliged to explain how the apparent further facts (for example mental facts) are reducible to, or supervenient on, facts at the primary level. This characterization of the issue about a reductionist thesis may be appropriate when the question is about materialism, since both materialists and dualists may agree about what the physical facts are (those that can be described in terms of the resources of physics – perhaps some extension or revision of present day physics). There may be some issues about exactly what the unproblematically physical

facts are, but the disputes between materialists about the mind and their opponents do not depend in any obvious way on how the level of physical fact is understood. But this way of characterizing a dispute between a reductionist and an anti-reductionist requires the presupposition that a category of fact to which all facts either are or are not reducible has been identified. In the case of Lewis's more abstract and general reductionist program, this presupposition cannot be taken for granted. One may object to the thesis of Humean supervenience, not because one thinks that there are causal facts that go beyond the facts about the local qualities of point-sized things, but because one thinks that we have no notion of local quality that does not already have a causal dimension. On one influential account, properties in general are causal powers (see Shoemaker 1980). Such an account will imply that one cannot even begin to describe the basic facts that characterize the world, or any possible world, without making use of notions that the Humean requires be reduced to something more basic. Lewis recognized the need to respond to this kind of objection, and his metaphysical theory addresses it explicitly. I will consider some of the consequences of the theory, and then close by considering how the problem of counterfactuals, and the more general problem of natural necessity, looks if one rejects the metaphysical picture, and the project of Humean reduction.

26.5 Natural Properties

Properties, for Lewis, are just sets. They are sets that may include merely possible things, which, according to Lewis's modal realism, are as real as any actual thing, so the identification of properties with sets is not stuck with the untenable consequence that properties that are coextensive *in the actual world* are thereby identical. Every set of possible individuals is a property, in the abundant sense of "property," but some properties are more natural than others, and some are perfectly natural. What is it for a property to be perfectly natural? Naturalness is of course itself a property – a property of properties – and so it is also a set – a set of sets. The property of being natural is one of the primitives of the metaphysical theory, to be explained in terms of the philosophical work that it does. Natural properties are assumed to be intrinsic, and it is assumed that there enough natural properties so that one could give a complete specification of a possible world by specifying a spatiotemporal structure, locating a subset of the possible individuals in that structure, and saying what perfectly natural properties each of the individuals has (which natural sets it is a member of). Naturalness is a primitive notion, but we can get some grip on it by the examples of properties that Lewis takes to be paradigm cases of perfectly natural properties: the properties that our most fundamental science identifies are the perfectly natural properties that are exemplified in our world. These properties will be exemplified in some other possible worlds as well, but there will also be natural properties that are alien to our world, which means that they are unexemplified there.

Might the natural properties be dispositional? Might their nomological roles be essential to them? Lewis has an argument that the answer should be no. In fact, he goes so far as to claim that there is a possible world in which very different natural properties (a quark color and a quark flavor, in his example) trade places. "The two possibilities are isomorphic, yet different" (Lewis 1986b, 162). Here is the argument against the hypothesis that the nomological role of a property is essential to it:

> Start with a world where the quark colours and flavours do figure in the laws that are supposed to be essential to them. By patching together duplicates of things from that world, we can presumably describe a world where those laws are broken; yet perfectly natural properties are intrinsic *ex officio*, and so they never can differ between duplicates. The principle of recombination seems to me very compelling indeed. (Lewis 1986b, 163)

Duplicates, for Lewis, are individuals that share all intrinsic properties. The principle of recombination is (roughly) the principle that if there is a possible world containing an individual *a* and a possible world containing an individual *b*, then there is a single possible world containing both a duplicate of *a* and a duplicate of *b*. The basic idea of the argument if the principle of recombination is true, then no intrinsic property can have its nomological role essentially.

As a response to a critic of Humean supervenience, this argument has little dialectical force, since the critic has no commitment to a principle of recombination, or to the *ex officio* requirement that the most fundamental properties of things are intrinsic. Further, the denier of Humean supervenience may question the assumption (implicit in the order of explanation in Lewis's reductionist project) that laws of nature are the most basic notion in the natural-necessity family – that causal powers and dispositional properties are best explained in terms of laws, which are a kind of global regularity. For the Humean, causation is a not local relation, but depends constitutively on patterns of events throughout history, and so the causal powers, capacities, and dispositions of a thing will also depend on global patterns of the behavior of that thing and others – patterns that are described by the laws. But the anti-Humean rejects this assumption, taking causation to be a real relation, and this allows that the causal powers of an object may be intrinsic to it. Laws are presumably generalizations about causal facts and causal powers, but the anti-Humean rejects the assumption that causal powers are derivative from the generalizations about them, and so rejects the assumption that his claim that a certain property is essentially causal or dispositional can be put as a claim about nomological role.

The upshot is that Lewis's argument that natural properties are independent of their causal or nomological role presupposes Lewis's Humean metaphysical theory, and is not an argument for it. This does not deprive the argument of interest; I think it is best understood as an argument that helps to bring out some of the consequences of the metaphysical picture. But if the consequences are implausible, the argument may help to motivate a rejection of the premises that have these consequences.

The argument brings out that the metaphysical theory underlying Lewis's Humean project seems to require a radical gulf between the essential nature of a fundamental property – its quiddity – and the ways that the property manifests itself in the world. Lewis is up front about this commitment: as we noted, he holds that there is a possible world like ours except that a certain quark flavor and a certain quark color trade places. That is, in this possible world, things with the color behave exactly the way things with the flavor behave in our world, and vice versa. The distribution of the colors is exactly like the distribution of flavors in our world, and vice versa. Let's look more closely at the conception of natural property that this seems to require.

This alleged possibility of interchange of properties calls to mind Goodman's notorious predicates, "grue" and "bleen": an object is grue at any time t iff it is green at t and first examined before some fixed time (in the future), T, or else blue at t and first examined after T. "bleen" is defined in a similar way, with "green" and "blue" interchanged. (See Goodman 1983, chapter 3.) Let me pursue the analogy. If the interchange Lewis considers is possible, then so is a world where the quark color and the flavor (just for definiteness, let's say they are *green* and *strange*) trade places within the world at a certain fixed time. So let our gruish (or perhaps grangish) world be one where at midnight GMT, January 1, 2000, these two very different fundamental properties trade places. *Green* and *strange* each take on the nomological role of the other, and all the green quarks suddenly become strange, and the strange quarks become green. (I don't understand any of this, which probably helps, but if these natural properties are as detached from their roles as Lewis claims, then this supposition should be unproblematic.) Perhaps this interchange actually happened in our world – who could tell? Maybe there was a catastrophic millennial bug after all, but no one noticed. If this change in fact happened, then for the past decade and more physicists have been systematically

confusing greenness and strangeness. The true laws, which are those of the best systematization of the true natural properties, will be laws that mark the dramatic switch at the start of AD 2000, but we (or the people in the world I am describing) will never discover these laws. Now remember that properties are nothing but sets – it is just that only some sets are *natural* properties. Consider not the property *green* itself, but instead the nomological role of *green* (now assuming that the dramatic switch did not in fact take place, so that the actual nomological role of the quark color is actually much as physicists take it to be). The nomological role property has the same extension in our world as the property *green*, but in other possible worlds, such as the grangish world I described, they have different extensions. Still, the nomological role property is a perfectly good property (in the abundant sense of property). Why is it less natural? One might think that a small change in Lewis's metaphysical could allow it to accommodate the thesis that natural properties are causal powers. Just say that a slightly different set of properties is the set of natural ones: the nomological role properties. But if we took the natural properties to be defined by nomological role, then even granting the primacy of laws in the order of analysis, and the best systems account of laws, this would compromise the Humean reduction, since the best system of laws and the choice of natural properties in terms of which the worlds are described would be interdependent. And as Lewis's argument shows, it would also cast doubt on a principle of recombination, or on the assumption that natural properties must be intrinsic.

The Humean project (not Lewis's but Hume's, and that of the twentieth-century empiricists, such as Goodman) was originally motivated by empiricist epistemological principles. Empiricists in this tradition took regularities to be less problematic than unobservable causal processes and underlying "occult" powers because generalizations about the phenomena, even if they faced a problem if induction, were on the surface, and were at least falsifiable. If I have interpreted Lewis correctly, then his version of that project leaves that motivation behind, and in fact requires a radical disconnect between the metaphysical account of reality and the epistemological account of how we know about it. If the natural properties that are the referents of our terms are as isolated from the laws by which those properties manifest themselves as Lewis's metaphysical picture seems to imply, then it is hard to understand how we ever know what the laws are, or how the properties of things connect with how the things with those properties behave. Even if a satisfactory epistemology could be combined with Lewis's metaphysical picture, it is no longer obvious what motivates the Humean strictures. Why, if not for the empiricist's reasons, is the family of concepts involving some kind of natural necessity problematic? I don't want to suggest that these concepts are not problematic, but if we give up on the project of Humean reduction, then we need to say more about what the problem is. I will conclude, in section 26.6, with a brief remark about this.

26.6 Conclusion

In motivating the analysis of counterfactuals in terms of what is true in "close" possible worlds, Lewis considers the following challenge: "It's the character of our world that makes some A-worlds be closer to it than others. So after all it is the character of our world that makes the counterfactual true – in which case, why bring the other worlds into the story at all?" The reply is that "it is only by bringing other worlds into the story that we can say in any concise way what character it takes to make the counterfactual true. The other worlds provide a frame of reference whereby we can characterize our world" (Lewis 1986b, 22). For the defender of Humean supervenience, the frame of reference allows us to describe the actual world more concisely, and at a certain level of generality, but, for the anti-reductionist, the possible-worlds frame of reference is playing an

indispensable role in forming the concepts by which we describe our world at the most fundamental level. To describe the world just is to locate it in a space of possibilities, and one cannot separate the task of describing the world from the task of characterizing the space of worlds and the way our world is related to others. The development of a theory for predicting and explaining the phenomena involves the formation of the concepts by which the phenomena are described, and the testing of a theory against the evidence is also the testing of the legitimacy of the concepts which are formed.

The constructive task that remains if one replaces the Humean reductionist project with this more holistic picture is to develop a substantive account of the way that reference to other possibilities, and to a structure of relations between possibilities, contributes to fixing the meaning and reference of the terms we use to describe and explain the phenomena. This can be expected to involve connecting the inductive principles by which we test hypotheses with the principles by which we form the concepts that are used to structure the space of possibilities that our hypotheses distinguish between. More generally, it can be expected to involve clarifying the relation between the objective modal concepts and analogous subjective or epistemic modal concepts (causal necessity and epistemic certainty, chance and credence, causal independence and epistemic irrelevance, subjunctive and indicative conditionals). The idea of making this kind of connection has its roots in Hume, who emphasized the central role of causation in inductive reasoning, and connected the problem of induction with the problem of causation. It also connects with Goodman's observation that the problem of disposition and the problem of induction are two aspects of a common problem. The conceptual projection involved in the forming of dispositional concepts exploits the same principles as the inductive projection involved in confirming theories. The abstract semantics for conditionals of the kind that David Lewis and I proposed for conditionals offers a framework for generalizing this kind of connection.

Notes

1 All Goodman quotations in this section are from Goodman (1983), with page references in parentheses. The first edition was published in 1953, and chapter 1 is a paper discussing the problem of analyzing counterfactuals, first published in 1946.

2 This is only roughly right because it doesn't cover the case where there are no closest ϕ-worlds to x because there is an infinite sequence of ϕ-worlds that are closer and closer to x, with no last term. In this case, the conditional will be true in x if and only ψ is true at some ϕ-world, y, and also at all ϕ-worlds that are at least as similar to x as y is.

3 The label, and my discussion in that paper, may be misleading because they blur the line between two different distinctions: (1) between abstract formal constraints on a relation of comparative similarity and substantive constraints on the relevant respects of similarity, and (2) between constraints that are part of the semantics of the conditional and constraints that are determined by the context of use of the conditional.

References

Fine, K. (1975). Review of Lewis, *Counterfactuals*. *Mind*, 84, 451–8.

Goodman, N. (1983). *Fact, fiction and forecast*. Cambridge, MA: Harvard University Press.

Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs* 13, 145–52. Reprinted (with postscripts) in Lewis (1986a), pp. 32–66. Page references are to the latter version.

Lewis, D. (1986a). *Philosophical papers*, Vol. II. Oxford: Oxford University Press.
Lewis, D. (1986b). *On the plurality of worlds*. Oxford: Blackwell.
Quine, W. (1960). *Word and object*. Cambridge, MA: MIT Press.
Shoemaker, S. (1980). Causality and properties. In P. van Inwagen (Ed.), *Time and cause*, pp. 109–35. Dordrecht: Reidel.
Stalnaker, R. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory*, pp. 98–112. Oxford: Blackwell.

27

On the Plurality of Lewis's Triviality Results

ALAN HÁJEK

27.1 Introduction

"Probabilities of conditionals are conditional probabilities." "Desires are beliefs about what is good." Offhand, these claims seem to have nothing to do with each other. A tenuous connection between them is that they posed potential threats to important philosophical views of David Lewis's, and that he provided what he and many other philosophers took to be formal refutations of them. A rather more significant connection is that these apparent refutations used the same distinctive and innovative technique. Lewis introduced a new kind of *reductio ad absurdum* style of argument: while the claims, suitably formalized, do not lead to outright contradiction, he showed they are tenable only in *trivial* ways. Thus, Lewis proved what are known as *triviality results* against these claims. In doing so, he pioneered two cottage industries of providing more and more refined triviality results against various formulations of the respective claims. I will argue in this chapter that the connections go considerably further still: the two claims give rise to two debates that display a remarkably parallel structure. I hope that bringing out their many similarities will illuminate both of them, and will point the way to some promising avenues for future research.

27.2 Probabilities of Conditionals as Conditional Probabilities

27.2.1 Background and Terminology

The thesis that probabilities of conditionals are conditional probabilities has at its heart the following equation:

(PCCP) $P(A \rightarrow B) = P(B \mid A)$ for all A and B, provided $P(A) > 0$

'$\rightarrow$' is a connective, which we will interpret as a conditional, and P is a probability function. (We assume the usual formula for conditional probability,

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

$$P(B|A) = P(AB)/P(A), \text{provided } P(A) > 0$$

and that all requisite probabilities are defined.) Suitably quantifying and qualifying the equation gives us versions of 'Stalnaker's Thesis', so called since Stalnaker (1970) was an early proponent of it – although he later disavowed it, and indeed provided a powerful argument of his own against it (1976). But the thesis has the support of intuition, it has shaped much subsequent discussion about conditionals, and it lives on in various forms; Stalnaker deserves much of the credit for this.

Suitably quantifying and qualifying – but how? There are two free variables in PCCP, '→' and '*P*', over which we must quantify. → assigns to each pair of propositions $<A, B>$ a proposition $A \rightarrow B$, but there are many possible such →. First, let's consider both orders of the quantifiers:

Fixed → Universal version: There is some → such that for all *P*, PCCP holds.
Indexical → Universal version: For each *P* there is some → such that PCCP holds.

Now, let's consider a natural restriction on the *P* over which we quantify:

Fixed → for Rational Agents version: There is some → such that for all *P* that could represent a rational agent's credences, PCCP holds.
Indexical → for Rational Agents version: For each *P* that could represent a rational agent's credences, there is some → such that PCCP holds.

Each of these versions has some claim to being called 'Stalnaker's Thesis' (the literature typically conflates them), although I think the Fixed → for Rational Agents version formulates his (1970) position best.

If PCCP holds for the pair $<\rightarrow, P>$, we will say that → is a *probability conditional for P*. If → is a probability conditional for every probability function in some class of probability functions, we will call → a *probability conditional for* the class. Say that *P* is a *trivial* probability function if it has at most four values.

P_X is derived from *P* by *conditioning* on *X* if, for all *A* and *X*,

$$P_X(A) = P(A|X), \text{provided } P(X) > 0$$

P_x is derived from *P* by *two-celled Jeffrey conditioning* if there is a *C* such that

$$P_x(B) = P_x(C)P(B|C) + P_x(\neg C)P(B|\neg C)$$

where $P_x(C) = P(C) + x$. *Jeffrey conditioning* in general extends this rule to the cells $\{C_1, C_2, \ldots, C_n\}$ of a finite partition.[1]

We will typically regard *P* as the credence function of a rational agent. She also a *value* function *V*, which assigns to each proposition a real number representing how desirable it is to her. According to Jeffrey's (1983) *evidential decision theory*, *V* obeys the following additivity rule. If $\{Z_i\}$ is a partition, then

$$V(A) = \sum_i V(AZ_i)P(Z_i|A)$$

This is also called *the expected utility* or *expectation* of A, though it would be more accurate to call it the *conditional expectation*, since the associated weights are probabilities *conditional* on A. I will return to this thought at the very end of the chapter.

27.2.2 Why Stalnaker's Thesis Mattered to Lewis

Lewis opens his seminal (1976) article with a puzzle concerning conditionals. Ordinarily, assertability goes by probability. How assertable you should regard a sentence to be is measured by your credence in that sentence: the assertability of A is $P(A)$.[2] But indicative conditionals appear to be exceptions to this pattern. The assertability of the indicative conditional $A \rightarrow C$ seems to go instead by the conditional probability $P(C \mid A)$ – this has come to be known as *Adams's Thesis* (Adams 1975). Lewis finds Adams's Thesis compelling, and in need of explanation. The best explanation would be that such conditionals also follow the pattern: the assertability of $A \rightarrow C$ *also* goes by $P(A \rightarrow C)$, as usual. But this equals $P(C \mid A)$, as per Stalnaker's Thesis.

And yet the thesis would spell trouble for other Lewisian doctrines concerning conditionals. Lewis believed that the truth conditions for the indicative conditional are those of the material conditional, and that the probability of a conditional is the probability of its truth. But the thesis fails for the material conditional. The material conditional 'if the die lands odd, it lands 5' is true if either the die does not land odd or it lands 5; this has probability 4/6. But the conditional probability P(5 | odd) = 1/3. Indeed, it is easy to show that the probability of a material conditional and the corresponding conditional probability agree only in trivial cases.

Nor can the thesis hold for the counterfactual, by Lewis's lights. In the early 1970s, he was engaged with Stalnaker in a classic debate over the truth conditions for counterfactuals. Their main point of dispute was conditional excluded middle, according to which

$$(A \rightarrow B) \vee (A \rightarrow \neg B)$$

is a tautology (with '→' now understood as a counterfactual connective). Stalnaker advocated conditional excluded middle, and indeed he made it an axiom of his logic of conditionals. Lewis opposed it. But it is easily shown that

$$P((A \rightarrow B) \vee (A \rightarrow \neg B)) = 1$$

for any function P that obeys PCCP. Yet, according to Lewis's semantics for the counterfactual, $(A \rightarrow B) \vee (A \rightarrow \neg B)$ could rationally be assigned probability less than 1. For example, according to Lewis, both 'if Bizet and Verdi had been compatriots, they would have been French' and 'if Bizet and Verdi had been compatriots, they would not have been French' are false, and so their disjunction is false, as a rational agent can determine.

Finally, important Lewisian views about decision theory are incompatible with the truth of Stalnaker's Thesis (although this concern did not arise until after Lewis's first salvo of triviality results). Jeffrey's expectation formula weights utilities by conditional probabilities of the form $P(Z_i \mid A)$ across a partition $\{Z_i\}$. Lewis believed that evidential decision theory gave wrong answers in Newcomb's Problem, and other such problems in which one's actions are evidentially relevant but causally irrelevant to the members of the partition. He advocated causal decision theory in its place, and, while he provided his own formulation of this theory, he emphasized that various alternative versions of the theory are reformulations of "one common idea" (Lewis 1981, 5). In particular, Gibbard and Harper's (1978) version replaces Jeffrey's conditional probability weights with the corresponding probabilities of counterfactuals, $P(A \mathrel{\square\!\!\rightarrow} Z_i)$. But if Stalnaker's Thesis is true for the counterfactual,

this replacement makes no difference at all. Gibbard and Harper's theory would then collapse into Jeffrey's theory, even in Newcomb-like problems – flying in the face of the "one common idea" that allegedly unites all versions of causal decision theory in opposition to Jeffrey's theory.

So a number of strands in Lewis's thinking were bound up with the fate of Stalnaker's Thesis. It was thus fitting that he played such an important role in determining its fate.

27.2.3 Why Believe Stalnaker's Thesis?

The strongest reason in favour of the thesis is surely that *it sounds right*. As Stalnaker writes, "The probability of the proposition, if Nixon is nominated then Johnson will win, should be the same as the probability that Johnson will win, on the condition that Nixon is nominated" (1970, 120), and so it apparently goes for the probability of any conditional.

Adams's Thesis gives us another reason to believe Stalnaker's Thesis: the former is apparently compelling, and its best explanation seems to be the latter.[3] The so-called 'Ramsey Test' provides further support. Ramsey (1965) suggests that you evaluate the conditional 'if A, then B' by hypothetically adding A to your system of beliefs, minimally revising what you currently believe in order to do so; then, evaluating B on the basis of your revised body of beliefs. $P(A \rightarrow B)$ measures how well the conditional performs on Ramsey's test. But apparently $P(B \mid A)$ does too. For conditioning on A seems to capture the notion of 'minimally revising what you currently believe in order to accommodate A'; and your evaluation of B in your new belief state $P(_ \mid A)$ is just $P(B \mid A)$.

So much for reasons to believe Stalnaker's Thesis. But Lewis provided weighty reasons to disbelieve it.

27.2.4 Lewis's Triviality Results against Stalnaker's Thesis

Lewis's triviality results have been collectively described as a "bombshell" by van Fraassen (1976), Stalnaker (1976), Harper (1981), Edgington (1995), and Bennett (2003) – this gives one a sense of their importance. He presents them in two papers, roughly a decade apart.

27.2.4.1 Lewis (1976) While Lewis presents two triviality results in this paper, logically speaking there is only one: the second one entails the first, and their proofs are almost identical – so much so that it is perhaps a little surprising that Lewis distinguishes them. So let us confine our attention to the stronger result.

> **Second triviality result:** If a class of probability functions is closed under conditioning, then there is no probability conditional for that class unless the class consists entirely of trivial probability functions.

Lewis conclusively refutes here the Fixed $\rightarrow$ Universal version of the thesis, and apparently also the Fixed $\rightarrow$ for Rational Agents version. After all, there are non-trivial rational credence functions, yet it seems that the class of rational credence functions is closed under conditioning – every case of conditioning apparently corresponds to a possible learning episode that some rational agent might have.

Lewis concludes that "The quest for a probability conditional is futile" (298). Is it?

27.2.4.2 Closure under conditioning Is the class of all rational credence functions really closed under conditioning? Starting with a rational credence function P, can conditioning yield a non-rational

credence function? It certainly yields a probability function; but surely some probability functions are not rational credence functions. Moore-paradoxical propositions such as 'it is raining and nobody believes it is raining' may warrant positive credence, but it seems they can never be (rationally) believed. Conditioning P on such propositions then arguably yields non-rational credence functions. It is no problem for the Fixed $\rightarrow$ for Rational Agents version, then, if there is no probability conditional for a class of probability functions that includes *these* functions.

But stronger triviality results are in the offing.

27.2.4.3 Lewis (1986) Ten years later, Lewis motivates his further triviality results with two related objections to his results thus far. Presumably one conditions on one's *total evidence*, and not all propositions are fit to be anyone's total evidence (e.g. Moore-paradoxical propositions). So arguably, the class of rational credence functions is *not* closed under conditioning; at best, it is closed under conditioning on evidence propositions.

Lewis replies that the set of evidence propositions for a subject at a given time form a partition; moreover, "the subject's limited powers of discrimination will ensure that this partition is a finite one" (1986, 583). He then appeals to his

> **Third triviality result:** If a class of probability functions is closed under conditioning on the propositions in some finite partition, then there is no probability conditional for that class unless the class consists entirely of trivial probability functions.

But there is another reason to question whether the class of rational credence functions is closed under conditioning. Indeed, it is a reason to insist that the class is, so to speak, *anti-closed* under conditioning: any instance of conditioning takes one *outside* the class! One might argue (à la Appiah 1985) that a rational credence function must be *regular* – it assigns probability 1 only to necessary truths. The thought is that assigning probability 1 to a proposition that could be false is a kind of overconfidence, which among other things will manifest itself in irrational betting dispositions. Moreover, the result of conditioning on some contingent evidence E is irregular, since it assigns probability 1 to E.

Lewis responds with his

> **Fourth triviality result:** There is no probability conditional for any class of probability functions closed under two-celled Jeffrey conditioning, unless the class consists entirely of trivial functions.

It follows that there is no probability conditional for any class of probability functions closed under Jeffrey conditioning, unless the class consists entirely of trivial functions – for two-celled Jeffrey conditioning is a special case of Jeffrey conditioning.

What Lewis actually showed is considerably stronger than the result he advertises. His own proof really shows that if P is a certain 'parent' distribution for which $\rightarrow$ is a probability conditional, and P_1 and P_2 are two different distributions descended from P by two-celled Jeffrey conditioning, then $\rightarrow$ is a probability conditional for at most one of P_1 and P_2.

In response, one might question whether Jeffrey conditioning is an acceptable revision rule; if it is not, then we might doubt that the set of rational credence functions is closed under it. One complaint is that it is *non-commutative*: two successive Jeffrey conditionings may yield a different result from what one gets by reversing their order (see Levi 1967).

As well as spawning a new area of research, Lewis's triviality results have spawned a host of misreadings and misstatements of what he showed. To some extent, this is understandable, as the

results are somewhat unintuitive – in particular, their appeals to classes of probability functions closed under various operations – and they resist easy paraphrases. In fact, couching all his results in terms of 'closure under' some rule or other understates what he actually proved in each case, and soon we will see that his results can be considerably strengthened. They cannot be summarized with simple slogans like 'probabilities of conditionals cannot be conditional probabilities', or even 'any probability function for which the probability of a conditional equals the corresponding conditional probability is trivial'.

Indeed, Lewis is well aware that probabilities of conditionals *can* be non-trivially equated to conditional probabilities:

> Once we give up on universality [the Fixed → Universal version], it may be encouraging to find that probability conditionals for particular probability functions, at least, commonly do exist. Given a probability function P, we may be able to tailor the interpretation of → to fit. Suppose that for any A and C there is some B such that P(B/A) and P(C/A) are equal if both defined; this should be a safe assumption when P is a probability function rich enough to represent someone's system of beliefs. If for any A and C we arbitrarily choose such a B and let A → C be interpreted as equivalent to AC v ¬AB, then → is a probability conditional for P. (1986, 301)

However, Lewis believes that this loophole for Stalnaker's Thesis is philosophically untenable:

> But such piecemeal tailoring does not yet provide all that we want. Even if there is a probability conditional for each probability function in a class, it does not follow that there is one probability conditional for the entire class. Different members of the class might require different interpretations of → to make the probabilities of conditionals and the conditional probabilities come out equal. But presumably our indicative conditional has a fixed interpretation, the same for speakers with different beliefs, and for one speaker before and after a change in his beliefs. Else how are disagreements about a conditional possible, or changes of mind? (301)

This argument anticipates some important anti-contextualist arguments in other domains (e.g. MacFarlane (2005) regarding knowledge). The thought is that contextualist treatments of some piece of language threaten to make genuine disagreements and retractions look merely like instances of people talking past one another. (Compare: there is no real disagreement when you say "I'm hungry," and I say "*I'm* not hungry.") The problem seems to be especially pressing in this case, since conditionals on this view threaten to be acutely context-dependent, potentially shifting interpretation with each shift in speaker or believer.

It would suffice to evade Lewis's results if the → were not stable in interpretation throughout an entire class of probability functions closed under conditioning, or restricted conditioning, or Jeffrey conditioning – but this could still be achieved by an →that was uniform throughout *most* of such a class. This might prompt the hope that such a near-uniform → might do well as an approximation to a conditional of natural language, and still be a probability conditional for all rational credence functions.

However, we may scathe even this modest hope. Hájek and Hall (1994) strengthen Lewis's first three triviality results, showing that if → is a probability conditional for *P*, *every* conditioning yields a function for which → is not a probability conditional for *P* (trivial cases aside). Merely observing that some conditionings are not legitimate (for one reason or another) will not save the Fixed → for Rational Agents version. As well as being stronger, this is a rather more intuitive and easily statable result than Lewis's. Even uniformity of the → across just a pair of probability functions, related by conditioning, is uniformity enough to give the stronger result a toehold.

But perhaps that is still too much uniformity. Van Fraassen (1976, 274) responds that "our probabilities are inextricably involved in the way we represent the possibilities, and nearness relations among them, to ourselves . . . if our ideas about the one change, will we not revise our modelling of the other?" He apparently assumes that conditionals should be given a Stalnaker/Lewis-style 'nearest possible world' semantics. It appears that on van Fraassen's view the conditional is so intimately tied to opinion that *any* change in opinion entails corresponding changes in which propositions are picked out by conditional utterances.

Hájek (1994) provides a *perturbation* argument that drops the fixed interpretation of the → altogether. Roughly, the idea is that for any given A and B, *whatever* proposition $A \rightarrow B$ turns out to be, probability can be shifted so as to change one side of PCCP, while holding the other side constant (trivial cases aside). Hence, if PCCP holds initially, it is cleaved by this shift. Crucially, we do not need to assume that → has a fixed interpretation throughout these probability movements. If → behaves indexically, sensitive to the probability function in whose scope it appears, then at least some changes in the distribution of probability result in changes in the identity of $A \rightarrow B$, for some A and B. To be sure, *some* such changes may exactly counterbalance the deleterious effects of these probability movements. However, some do not counterbalance it enough, some do not counterbalance it at all, and some actually accentuate it. To uphold PCCP, we must assume not merely indexicality, but that the indexicality is of just the right sort.

Then let us assume exactly that, you may say. Van Fraassen (1976) has a striking positive result that shows that such indexicality *will* ensure the unassailability of a certain form of Stalnaker's Thesis, but as we will soon see, it is a highly restricted form.

27.2.5 Indexicality

Before we get to it, let us take stock. The Fixed → Universal and Fixed → for Rational Agents versions of Stalnaker's are, I think, decisively refuted. However, we have seen nothing so far that impacts upon the other two versions:

Indexical → version: For each P there is some → such that PCCP holds.
Indexical → for Rational Agents version: For each P that could represent a rational agent's credences, there is some → such that PCCP holds.

And these are the versions that van Fraassen seems to have in mind for rehabilitating Stalnaker's Thesis.

A result in Hájek (1989), which is strengthened in Hájek (2012), refutes the first of these, and casts serious doubt on the second. Any non-trivial finite-ranged probability function has more distinct conditional probability values than distinct unconditional probability values. This means that the function's conditional probabilities cannot all be matched with its unconditional probabilities. *A fortiori*, this means that its conditional probabilities cannot all be matched with its unconditional probabilities *of conditionals*. There will always be some conditional probability that finds no match among the unconditional probabilities, and this will be a counterexample to Stalnaker's Thesis: it will be a conditional probability of the form $P(B \mid A)$ that does not equal $P(A \rightarrow B)$ (or indeed anything of the form '$P(X)$'). It is like a wallflower at a dance, unable to find a partner.

This result refutes the Indexical → Universal version, and it casts considerable doubt on the Indexical → for Rational Agents version. For it is implausible that *every* rational credence function has infinite range (or at most four values), as is needed to sustain a probability conditional.

But suppose for the sake of the argument that all rational credence functions have such a range. Hall (1994) shows that if we give → just a modicum of logical structure, even this is not

enough to save the Indexical → for Rational Agents version. *P* needs to have an *uncountably* infinite range if it is to have a probability conditional that obeys modus ponens. Indeed, *P* needs to be *full*: for every proposition *A* to which *P* assigns probability, *P* takes every value in [0, *P*(*A*)] over the subsets of *A*. To be sure, the validity of modus ponens has been questioned (e.g. McGee 1985; Etlin 2009). But it is salutary that the proponent of the Indexical → for Rational Agents version who admits that some rational credence functions are not full now is *forced* to deny the validity of modus ponens.

That said, van Fraassen (1976) has a positive result that finally provides some good news for the friend of Stalnaker's Thesis. If *P is* full, then it *can* sustain PCCP for an → with a conditional-like logic (which he calls "CE"). And restricting the compounding of sentences with → allows still more logical strength (C2, Stalnaker's preferred logic). The proof is demanding, but the basic underlying idea is simple. If *P* is full, then whatever value a conditional probability $P(B \mid A)$ might take, there will be *some* proposition whose unconditional probability is that value. Van Fraassen's construction involves a series of judicious choices of such propositions, to be regarded as conditionals.

See Hájek and Hall (1994) for an extensive survey of all these and other results concerning the fluctuating fortunes of the variants of Stalnaker's Thesis.

27.2.6 No-Truth-Value Accounts of →

But perhaps conditionals are not propositions at all. Perhaps they do not have truth conditions, or truth values. Rather than stating that the world is one way or another, perhaps their purpose is something else. In keeping with Adams's Thesis, a plausible purpose for asserting 'if *A*, then *B*' is to convey that one's conditional probability for *B*, given *A*, is high. Adams speaks of conditionals having probabilities, but he did not regard these as probabilities *of truth*. Importantly, conditionals for Adams do not enter into Boolean combinations the way declarative sentences do. McGee (1989) shows that Adams's approach can be extended to some limited Boolean combinations, but they still do not figure in a Boolean algebra. As a result, their 'probabilities' are not probabilities in the standard sense (which *are* defined over a Boolean algebra).

These moves block the triviality results that we have seen, which assume that conditionals express propositions. And a key motivation for the no-truth-value account is the thought that the triviality results *must* be blocked, since Stalnaker's Thesis is so compelling to the ear (see e.g. Edgington 1995), at least for indicative conditionals. That said, we might question whether Adams and others entirely uphold it. After all, as Lewis writes, "if it be granted that the 'probabilities' of conditionals do not obey the standard laws, I do not see what is to be gained by insisting on calling them 'probabilities' " (1976, 304–5). And the thesis that some *other* attribute of conditionals – e.g. their "assertabilities" – equals their corresponding conditional probabilities is not so compelling to the ear.

Lewis has an "inconclusive objection" to the no-truth-value hypothesis: it "requires too much of a fresh start. It burdens us with too much work still to be done, and wastes too much that has been done already" (1976, 305). The work done that he has in mind is the compositional semantics we have for how truth conditions for compound sentences are determined by the truth conditions for their constituents. If conditionals don't have truth values, then we will need to come up with new rules for how the assertability conditions (or what have you) of compound sentences involving them are determined by the assertability/truth conditions of their constituents.

I would put the point slightly differently. Sometimes fresh starts are required in our theorizing, when our old ways of theorizing are inadequate; we should regard them as sunk costs, and move on. But the point is that the compositional semantics is itself compelling. Conditionals *do* embed naturally in various compound constructions, and we already have a good account of how the truth-values of

these constructions are determined by those of their parts. The no-truth-value account's version of the Frege–Geach problem is serious. And probabilities are normally probabilities *of truth* of their arguments; the default presumption should be that conditionals are no exception.

To be sure, the desire to uphold Stalnaker's Thesis mitigates this presumption somewhat. But perhaps we should question just how compelling it really is. In fact, philosophers who like nearest-possible-world accounts of conditionals ought to balk at it. $P(A \rightarrow B)$ should be sensitive to B's holding or not at the *nearest* A-worlds, whereas $P(B \mid A)$ pays no heed to that. And a change in the nearness ordering of worlds without a change in probabilities – for example, due to a change in context – may change $P(A \rightarrow B)$, but not $P(B \mid A)$.

In any case, the main support for the thesis comes from our mentally testing some instances (like Stalnaker's Nixon/Johnson instance), ticking them, and generalizing therefrom. However, Kaufmann (2004) has examples in which intuitive judgments conflict with the thesis. And we should be wary of generalizing too hastily from positive instances like this. Consider this thesis: *desirabilities of conditionals are conditional desirabilities*. Write this as

$$V(A \rightarrow B) = V(B|A)$$

where the right-hand side is understood as the desirability of B *given* A – the desirability of B restricted to worlds in which A is true. I submit that this thesis has about as much prima facie plausibility as Stalnaker's Thesis. Test some cases. The desirability of the conditional 'if it is sunny tomorrow, we will have a picnic' intuitively seems the same as the conditional desirability of our having a picnic, given that it is sunny tomorrow. Tick! Generalizing from instances such as this, we have the desirability thesis.

But the thesis is surely wrong. Groucho Marx famously said "I don't want to belong to any club that would have me as a member." For example, he may think that only a club with low admission standards would admit him, and he does not want to be a member of a club with such standards. Suppose that he highly values being a member of a particular club: V(he is a member) is high. But then he restricts his attention to worlds in which he *is* a member – for example, he learns that he has been made a member. Immediately the value of being a member drops: V(he is a member | he is a member) is *low*. This is *not* the value of the corresponding conditional, V(he is a member $\rightarrow$ he is a member). For that is the value of a tautology, something he regards with indifference.

This example prepares us for our transition to the second half of this chapter, which seemingly involves an entirely new topic from the probabilities of conditionals. But I hope to show along the way that our topics have much in common.

27.3 Desire as Belief

27.3.1 Background and Terminology

The thesis that desires are beliefs about what is good has at its heart the following equation:

(DAB) $V(A) = P(\mathring{A})$, for all A and B in the domain of V and of P, with $P(A) > 0$

'$\circ$' is an operator (read 'halo'), which we will interpret as attaching 'is good' to a proposition. (We assume that V is scaled to the [0, 1] interval.[4]) Suitably quantifying and qualifying the equation gives us the so-called "Desire-as-Belief Thesis," formulated and named by Lewis, although he gave powerful arguments against it.

Suitably quantifying and qualifying – but how? There are three free variables in DAB, 'V', 'P', and '^{o}', over which we must quantify, but there are many possible such ^{o}. We will regard $<V, P>$ as coming as a pair, since they greatly constrain each other, and soon we will interpret them as belonging to a single agent. First, let's consider both orders of the quantifiers: $P(A \rightarrow B) = P(B \mid A)$:

Fixed ^{o} Universal version: There is some ^{o} such that for all $<V, P>$, DAB holds.
Indexical ^{o} Universal version: For each $<V, P>$ there is some ^{o} such that DAB holds.

Now, let's consider a natural restriction on the $<V, P>$ over which we quantify:

Fixed ^{o} for Rational Agents version: There is some ^{o} such that for all $<V, P>$ that could represent a rational agent's desires and credences, DAB holds.
Indexical ^{o} for Rational Agents version: For each $<V, P>$ that could represent a rational agent's desires and credences, there is some ^{o} such that DAB holds.

Each of these versions has some claim to being called a 'Desire-as-Belief Thesis', but the Fixed ^{o} for Rational Agents version captures best the version that Lewis has in mind.

27.3.2 Why the Desire-as-Belief Thesis Mattered to Lewis

Lewis favored a broadly Humean view according to which there are no necessary connections between distinct existences. This view extends to the psychological states of beliefs and desires. Moreover, Lewis followed the Humean theory of motivation, that "reason is the slave of the passions," or, as he put it, "we are moved entirely by desire" (1988, 323), though I think that really he should say that desire and belief are equal partners in moving us. After all, another great influence on him, Ramsey, had shown how preferences could be represented by separate – and equally important – contributions due to desire and belief, or, better, degrees of desire and degrees of belief. Lewis was also inspired by the evidential decision theory of his long-time colleague Richard Jeffrey, which gave Ramsey's insight an elegant Bayesian formalization. In it, utilities and probabilities both earn their keep, working in tandem to determine or rationalize action. Lewis regarded decision theory as a formalization – idealized, to be sure – of folk psychology's commitment to the interaction of beliefs, desires, and actions, enshrined in his analytic functionalism.

If the Desire-as-Belief Thesis is true, then beliefs and desires are necessarily connected after all. Far from reason being slave to the passions, or even equal partners, it seems that reason has the driver's seat when it comes to motivation: desire is just belief with a particular content. Indeed, we could apparently rewrite decision theory solely in terms of credences, replacing all occurrences of 'V's of propositions with corresponding 'P's of their haloed counterparts. This, in turn, would be problematic for Lewis's analytic functionalism, for presumably the reduction of desires to beliefs is not part of folk psychology.

In "Desire as Belief II" (1996), Lewis's interest in the Desire-as-Belief Thesis is based more in metaethics. Reading DAB from right-to-left, we could have an account of what goodness consists in: it bears a certain relationship to being valued. Relatedly, according to Lewis (1989), something is a value for you iff you are disposed to desire to desire it (under suitably ideal conditions). But if Desire as Belief is true, A's being good would be a better candidate for A's being a value – necessarily, the attitude that one has to A's being good constrains the value one attaches to A.[5] Moreover, if valuing is a matter of having a belief that can be true or false, or credences that are better or worse (as measured by their distance from 1 for the truth), an objectivist rival to Lewis's theory looks promising.

So the Desire-as-Belief Thesis begins in Lewis's first paper as a psychological thesis about the relationship between beliefs and desires. So understood, there need be no connotation of *moral* goodness about *Å*; it simply corresponds to your degree of desire in *A* (where *A* might be your having an ice cream, or today's being sunny). In that case, we might just as well symbolize it as $A^{☺}$. But, in his second paper, the 'halo' has much more of an ethically charged reading.

27.3.3 Why Believe the Desire-as-Belief Thesis?

I doubt that the Desire-as-Belief Thesis has the prima facie plausibility that Stalnaker's Thesis does. Moreover, I find it a little curious how Lewis motivates it and formulates it.

Lewis canvases the broadly anti-Humean view that we can be moved by beliefs – to the extent that we are motivated by desires, these are *also* beliefs about what is good. He then introduces Frederic, "that famous slave of duty" (1988, 325), who he imagines is motivated entirely by desires that are necessarily connected to beliefs about what is good:

> Then the Desire-as-Belief Thesis says that Frederic desires things just when he believes they would be good. Or better, since we must acknowledge that desire and belief admit of degree, he desires things just to the extent that he believes they would be good. To any ordinary proposition A, there corresponds another proposition: Å, the proposition that it would be good that A. Frederic's expected value for A, which represents the degree to which he desires that A, equals the degree to which he believes that A.

Lewis codifies this with DAB. (Here and henceforth I replace his use of 'C' for the credence function with 'P' throughout for consistency with the way that we have presented Stalnaker's Thesis; bear this in mind when I quote him.)

He does not spell out the Desire-as-Belief Thesis here, or indeed anywhere in "Desire as Belief," in full generality. Frederic is a special case, one whom we might regard as pathological. At best, he is a placeholder for all slaves of duty (famous or otherwise) – hardly the anti-Humean's poster child. It is also unclear that DAB is the right codification of the preceding gloss. "The extent that he believes things would be good" admits of degree, but it is unclear whether the degrees attach to "he believes" (as Lewis has it), or rather to "good." In the latter case, the schema should be more:

$V(A) = \text{x iff Believe}(A \text{ is good to degree x})$

DAB collapses goodness into just two degrees, on or off. In so doing, it appears to be straightforwardly refuted by 'cross-over' cases, in which a proposition that *extremely probably* is *a little bit* good is compared with another proposition that *less probably* is *extremely good*. For example, it is extremely probable that helping an old lady is good: *P*(helping old lady is good) is close to 1; but it is good to a rather small degree. On the other hand, electing as president Ms. Honourable, a woman of great integrity and vision, is less probably good – the power of the high office may corrupt her, or her vision may be thwarted in various ways. *P*(Ms. Honorable's election is good) is not as close to 1. But her election is likely to be good to an extremely high degree, and, given the high stakes involved, overall we regard her election as far more desirable than helping the old lady. If Fredric is unable to make such assessments, because he has such impoverished evaluations of goodness, then he really *is* pathological. Why, then, are lessons learned about him of any consequence to anti-Humeans?

Lewis defends his strategy of focusing on the case of Frederic: "It is fair to take a simple case; because if our Anti-Humean's thesis collides with Decision Theory only in simple cases, that is bad enough" (1988, 325). It is true that if Frederic is a rational agent, and the anti-Humean thesis fails for him, then it cannot hold for all rational agents. However, this might raise the suspicion that

Lewis's formulation of the thesis is a straw man. It is hardly a strike against anti-Humeanism if the alleged trouble arises only for a "simple" (a euphemism for pathological?) case like Frederic's.

But it doesn't. Fortunately, later in "Desire as Belief," Lewis removes the simplifications of this case in two respects. First, he allows value to have two components: one that obeys DAB, and one that does not, and he argues this makes no real difference to his argument. Second, he formulates a version of the Desire-as-Belief Thesis that acknowledges various degrees of goodness, and that is thus far more plausible for agents in general, and his argument goes through again, *mutatis mutandis*; more on that shortly. So I think that the real rationale behind Lewis's strategy is that the same technique that he uses to refute Desire as Belief generalizes to refute the more plausible versions. Let us look, then, at that technique.

27.3.4 Lewis's Triviality Results against the Desire-as-Belief Thesis

Lewis presents his triviality results in two papers, roughly a decade apart.

27.3.4.1 Lewis (1988) Lewis (1988) presents his first result against the Desire-as-Belief Thesis; he does not explicitly call it a "triviality result," but we may justly do so. Begin with a proposition A and a $<V, P>$ pair for which DAB holds. He shows that apart from trivial cases, we can update by Jeffrey conditioning to a new pair $<V', P'>$ for which DAB no longer holds. The proof is algebraic and rather sophisticated: the probability of $\mathring{A}$ responds to the Jeffrey shift according to a certain linear function, whereas the desirability of A responds according to a quotient of such functions. Thus, the shift breaks the equality between the desirability of A and the probability of $\mathring{A}$.

27.3.4.2 Lewis (1996) Eight years later, Lewis calls this argument "needlessly complicated" (1996, 308), and he offers a simpler argument that does not rely on Jeffrey conditioning. He begins with this observation:

> DAB can be equivalently restated as a pair of equations: necessarily, for any A and P,
>
> (DACB) $V(A) = P(\mathring{A} \mid A)$
>
> (IND) $P(\mathring{A} \mid A) = P(\mathring{A})$.
>
> To derive DACB, we recall that DAB is supposed to continue to hold under redistributions of credence, and we redistribute by conditionalizing on A. (308–9)

He then argues that (IND) is the culprit: given certain assumptions about P(A) and $P(\mathring{A} \mid A)$, various redistributions of credence will make (IND) go from true to false. For example, moving probability from $\neg\mathring{A}$ & $\neg A$ to $\mathring{A}$ & $\neg A$ will increase $P(\mathring{A})$ while keeping $P(\mathring{A} \mid A)$ constant (Figure 27.1):

He could have stated the result in a way that strikingly parallels his triviality results against Stalnaker's Thesis: IND is not contradictory, but it holds only for trivial probability functions. Both Stalnaker's Thesis and the Desire-as-Belief Thesis imply independence conditions that only trivial functions satisfy.

This is all well and good; but remember the concern that Desire-as Belief is a straw man – at best plausible for agents like Frederic who distinguish only two levels of goodness, and perhaps not even such agents. So it is just as well, and good, that Lewis goes on in the first paper to generalize DAB for an agent who distinguishes multiple degrees of goodness, $g_1, \ldots, g_m$. Let $\mathring{A}_j$ be the proposition that it

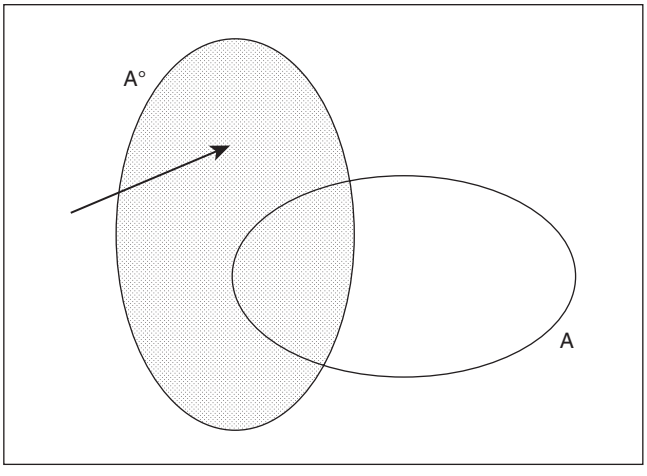

Figure 27.1 Moving probability.

would be good to degree g_j that A. Then what I'll call the Generalized Desire-as-Belief Thesis replaces (DAB) with:

(GDAB) $V(A) = \Sigma_j P(\mathring{A}_j)g_j$

Using an algebraic and rather sophisticated proof similar to before, Lewis shows that this entails the absurdity that the agent cannot both change her degree of belief and her degree of desire in A. As with Stalnaker's Thesis, the absurdity is not one of contradiction, but of triviality – there are $<V, P>$ pairs that meet this condition, but they are trivial.

Lewis concludes that the Desire-as-Belief Thesis "is in bad trouble" (1996, 329). Is it?

27.3.4.3 Invariance Jeffrey conditioning involves a redistribution of credences across the cells of a partition $\{E_1, E_2, \ldots E_n\}$. This will often result in a change in the expected value of a proposition A that is compatible with more than one of these cells. After all, the values of its conjunctions with the E_i, $V(AE_i)$, will often vary; the redistribution will often put more weight on the lower-value conjunctions, in which case $V(A)$ will decrease, or on the higher-value conjunctions, in which case $V(A)$ will increase. But Lewis maintains that if A is compatible with only one of the cells, then its value does not change. He calls this the *Invariance Assumption*, which we can formulate as:

$V_{new}(A) = V(A)$ if A is compatible with only one of the E_i

This is because, with Jeffrey conditioning, the internal profile of probability *within* any given cell stays rigid – that profile is merely uniformly pumped up or pumped down by the multiplication of a constant, its overall shape remaining the same. So the weightings given to the higher- and lower-value parts of that cell remain the same throughout the overall shift in probabilities.

The Invariance Assumption is not itself part of Jeffrey's decision theory. For all that theory tells us, utilities of highly specific propositions within the cells could spontaneously change. The additivity formula for expected value tells us how the expected value of a less specific proposition is

compounded from values of more specific propositions that entail it, but it is silent about what happens to *them*.

Nevertheless, Lewis argues that discarding the Invariance assumption has "an unintelligible consequence" (1996, 331). His argument begins: "We note first that it is impossible to discard Invariance only as applied to the AE_j's themselves. For each AE_j may be further partitioned into subcases $AE_jF_1, \ldots,$ AE_jF_k, in such a way that each of these subcases is *maximally specific in all respects relevant to its value*" (332, my emphasis). The idea is that since the profile of probability stays rigid within AE_j, any change in its value would have to issue from a change in value of at least some of its subcases – which is to discard Invariance for such subcases. But, Lewis continues, it is unintelligible to change the value attached to a proposition that is maximally specific in all respects relevant to its value. For that very specificity ensures that the proposition settles all questions of what would be good. So the value that one attaches to that proposition is settled, and thus cannot change.

I question this argument. Jeffrey's decision theory assumes an *atomless* Boolean algebra: every proposition X can be split into two more specific propositions X_1 and X_2 whose disjunction is X – ad infinitum. All of Jeffrey's propositions are *gunky*. It is far from obvious, then, that we will always be able to partition any proposition into subcases that are maximally specific in all respects relevant to value. Perhaps value is gunky too. Perhaps any proposition has within it subcases differing in value.

It may seem just as well, then, that Lewis goes on in "Desire as Belief II" to give another, simpler argument against Desire-as-Belief. But a version of Invariance is assumed there, too. Look again at his explanation of why DAB is equivalent to the conjunction of DACB and IND: "To derive DACB, we recall that DAB is supposed to continue to hold under redistributions of credence, and we redistribute by conditionalizing on A." This may seem a little fast. To be sure, when we redistribute by conditionalizing on A, the *right-hand side* of DAB becomes $P(\mathring{A} \mid A)$. But what about the left-hand side? Shouldn't it be updated too, to $V(A \mid A)$? That is, what we get from conditionalizing both sides of DAB is not immediately DACB, but rather:

$$V(A \mid A) = P(\mathring{A} \mid A)^6$$

Enter Invariance! We can derive DACB if we can assume

$$V(A \mid A) = V(A) \text{ (Invariance for Conditioning)}$$

Much as strict conditioning can be thought of as a limiting case of Jeffrey conditioning, Invariance for Conditioning can be thought of as a special case of Invariance. Consider the two-celled partition $\{A, \neg A\}$, and pump up the probability of A to 1, while keeping rigid the profile of probability within it. Now, A is compatible with only one cell of this partition (itself), so by Invariance, extended to this limiting case,

$$V_{\text{new}}(A) = V(A)$$

where

$$V_{\text{new}}(A) = V(A \mid A)$$

So I take the core of Lewis's argument to be:

1. DAB and Invariance for Conditioning jointly entail IND.
2. Invariance for Conditioning holds (being a limiting case of Invariance).

3. IND is untenable (leads to triviality).
∴ C. DAB is untenable (leads to triviality).

DACB is just a step in deriving 1.[7]

I have already raised a concern (atomlessness) about Lewis's argument for Invariance. But this concern is not an argument *against* Invariance, still less an argument against Invariance for Conditioning. For all that has been said, that assumption may be sound, and Lewis's refutation goes through. So let us challenge Invariance for Conditioning head on. Groucho Marx's "I don't want to belong to any club that would have me as a member" seems to provide a counterexample. As we saw, for him V(I am a member) was high, but V(I am a member | I am a member) was low. Byrne and Hájek (1997), Bradley and List (2009), and Kotzen (n.d.) have other examples of this kind.

Invariance for Conditioning, then, cannot be a constraint on *all* rational agents. But it is enough that *some* rational agents obey it, for whatever reason. Restrict our attention to such agents. For these agents, Lewis's argument goes through. It is still absurd that *they* must satisfy IND, and thus must have trivial probability functions. So I think Lewis's argument is still basically right.

Recall our discussion of how Lewis's triviality results against Stalnaker's Thesis assumed that the $\rightarrow$ has a fixed interpretation across different probability functions. Lewis does not address the issue of whether he assumes ^{o} has a fixed interpretation across different $<V, P>$ pairs, but he apparently does. It turns out that we can drop this assumption and Invariance. Paralleling my perturbation argument against Stalnaker's Thesis, we can show that for any given A, probability can be shifted so as to change one side of DAB or GDAB, while keeping the other side constant (trivial cases aside). Hence, if DAB or GDAB holds initially, it is cleaved by this shift. If ^{o} behaves indexically, sensitive to the probability function in whose scope it appears, then at least some changes in the distribution of probability result in changes in the identity of ^{o}, for some A. To be sure, *some* such changes may exactly counterbalance the deleterious effects of these probability movements. However, some do not counterbalance it enough, some do not counterbalance it at all, and some actually accentuate it. To uphold DAB or GDAB, we must assume not merely indexicality, but that the indexicality is of just the right sort.

Then let us assume exactly that, you may say.

27.3.5 Indexicality

Let us take stock. The Fixed ^{o} Universal and Fixed ^{o} for Rational Agents versions of Desire as Belief are, I think, decisively refuted. However, we have seen nothing so far that impacts upon the other two versions:

> **Indexical ^{o} Universal version:** For each $<V, P>$ there is some ^{o} such that DAB holds.
> **Indexical ^{o} for Rational Agents version:** For each $<V, P>$ that could represent a rational agent's desires/credences, there is some ^{o} such that DAB holds.

If we allow the identity of $\mathring{A}$ to change in just the right way as the distribution of probability changes in any of the arguments above, then for all we have seen so far, DAB may be upheld throughout.

Lewis (1996) is alert to this challenge, labeling such an indexicality thesis "Inconstancy." It recalls his observation that probability conditionals can be tailored to particular probability functions. There he responded that the conditional must have a fixed interpretation – else disagreements and retractions involving conditionals would not be possible. He might have responded similarly here: understanding the ^{o} as denoting the goodness of a proposition, it must have a fixed interpretation – else disagreements and retractions involving goodness would not be possible.

Instead, he points out that Indexical Desire as Belief will almost trivially hold – all it takes is for the range of V to be a subset of the range of C. Then, for any A, there exists a B such that $V(A) = P(B)$. He then dismisses this fact.

> It's too easy, and no anti-Humean should celebrate such an easy victory. The DAB equation holds not in virtue of any interesting relationship between a desired and a believed proposition, but only in virtue of what it takes for a proposition to deserve the name "Å" – and what it takes is nothing more or less than the right credence, one that equals V(Å). There is nothing at all anti-Humean about this little trick. Further, there is nothing that should make us want to say that Å is the proposition that A is objectively good. (1996, 312)

To be sure, if there are no constraints imposed by A on what can count as $\mathring{A}$, then this victory is indeed too easy. But Lewis has not shown that the victory cannot also be hard won. For all we have seen, $\mathring{A}$ might be determined by A and $<V, P>$ in a principled way, so that the halo over A is genuinely earned. In fact, for all we have seen it might even deserve the name 'objective goodness' – not that this required in order to uphold anti-Humeanism about desire.

Note that we have moved some distance from the original terms of the debate, which was simply whether beliefs are necessarily connected to desires. In fact, Lewis moves even further, suggesting in a footnote (1996, 312) that the real anti-Humean goal is to deliver *objective ethics*. Hájek and Pettit (2004) argue that even that goal has not been conclusively thwarted by Lewis's (and other) results. They maintain that even objective ethical theories can be indexicalist, citing as examples some versions of consequentialism, and non-consequentialist theories of rightness.

27.3.6 Reformulating the Anti-Humean Thesis

Lewis emphasizes that in "Desire as Belief" he refutes "one possible version of anti-Humeanism" (1996, 309). This might give the anti-Humean hope that other versions may be tenable. We have already seen one route that the anti-Humean might try: varying the quantifier structure over the DAB equation. Now let's consider modifying the equation itself.

Price (1989) argues that its right-hand side should be replaced with '*conditional* belief': $P(\mathring{A} \mid A)$. We have met the resulting equation before:

(DACB) $V(A) = P(\mathring{A} \mid A)$

The combination of DACB and IND, we saw, was equivalent to DAB; but the trivialization of DAB was due to IND, not DACB, so the latter might capture whatever is right in DAB. And Price argues that "whenever it makes a difference, we should assess a possible outcome under the hypothesis that it is the actual outcome" (1989, 122). So the proposal is to replace Desire as Belief with Desire as Conditional Belief, replacing DAB with DACB.

In his (1996), Lewis shows that Desire as Conditional Belief is equivalent to another version of an anti-Humean thesis, Desire by Necessity, which he introduces thus:

> Desire by Necessity is a comparatively simple and unproblematic version. In its simplest form, it says that necessarily and regardless of one's credence distribution, certain point-values must be high and the rest low. Scale these as 1 and 0. Let G be the union of point-propositions with necessarily high value . . . Then for any proposition A and any credence distribution P (provided that P(A) is positive . . .),

$V(\mathrm{A}) = \mathrm{P}(\mathrm{G} \mid \mathrm{A})$.

Point-propositions are "the propositions corresponding to the maximally specific possibilities" (1996, 303), and their V-values are *point-values*. Again, I question whether there are such things in Jeffrey's atomless decision theory, but let that pass. Importantly, Lewis regards Desire as Conditional Belief's equivalence with Desire by Necessity, as a *reductio:* "Desire as Conditional Belief is now unmasked" (312).

So what, you may ask, is wrong with Desire by Necessity? After all, Lewis himself says that it is a "comparatively simple and unproblematic version," and this is his entire evaluation of it – this sounds like *praise*! Immediately after the passage that introduces it, quoted above, he very briefly presents three refinements of it, with no further commentary. And then the section abruptly ends, and the topic changes. I am still left wondering: *what is wrong with it?* This for me is the most perplexing moment in the complete Lewisian corpus – so much so that it's as if some material went missing. And Desire as Conditional Belief is similarly left hanging – equivalent to Desire as Necessity it may be, but *what is wrong with that?*

At the very end of the paper, Lewis concludes telegraphically by saying that Desire by Necessity is "a form of anti-Humeanism, sure enough, but not the right form of anti-Humeanism" (313); and that's it. I do not know what makes a form of anti-Humeanism "right" or not, and why the form that says that we necessarily desire certain things – as it might be, those that conduce to The Good – is not. There is a useful paper waiting to be written about what Lewis might have had in mind here.

27.4 Some Future Avenues of Research?

I hope that I have displayed various structural similarities between the fluctuating fortunes of Stalnaker's Thesis, and those of Desire-as-Belief. Looking for parallels between their associated debates can provide guidance for further avenues for research. Lining up the various moves and countermoves side by side, as it were, we find some gaps: promising developments on one side that have no counterparts on the other. These developments can then be mimicked on the other side. They would be harder to discover by just looking at each debate in isolation.

27.4.1 The Desire as Belief Debate Informing the Stalnaker's Thesis Debate

Let's start with the last move that we have considered for rehabilitating the anti-Humean position: replacing the unconditional probability in DAB with a conditional probability in (DACB). The counterpart move on the side of Stalnaker's Thesis would be to replace the unconditional probability in PCCP with a conditional probability. As a result, we identify *conditional* probabilities of conditionals with conditional probabilities:

(CPCCP) $P(A \rightarrow B \,|\, A) = P(B \,|\, A)$, if $P(A) > 0$

We can then quantify over it as before, starting with the Fixed → Universal version. Like the original Thesis, this 'sounds right': the probability that if Nixon is nominated then Johnson will win, *on the condition that Nixon is nominated*, should be the same as the probability that Johnson will win, on the condition that Nixon is nominated. Moreover, far from being susceptible to triviality results, this thesis is trivially shown to be correct, given two assumptions about the logic of →. The left-hand side is

$$\frac{P((A \rightarrow B) \,\&\, A)}{P(A)}$$

Suppose only that → obeys modus ponens and centering (A & B entails $A \rightarrow B$). Then $(A \rightarrow B)$ & A is equivalent to A & B, so their probabilities are the same. Then this fraction immediately reduces to $P(B \mid A)$.

We could immediately derive PCCP if we could assume that conditionals are independent of their antecedents:

$$P(A \rightarrow B \mid A) = P(A \rightarrow B)$$

But like (IND) in Lewis's trivialization of DAB, this is the culprit. Perhaps part of the appeal of Stalnaker's Thesis is that it is natural to think of conditionals as independent of their antecedents – the truth or otherwise of 'if A, then B' prima facie does not seem to depend on whether A is true. The apparent demise of the thesis teaches us otherwise. It would be fruitful to explore how much of the bounty of the thesis, and of Adams's Thesis, can be recovered by CPCCP.

27.4.2 The Stalnaker's Thesis Debate Informing the Desire-as-Belief Debate

As far as I am aware, the Indexical ^{o} for Rational Agents version is still alive. A 'wallflower' result similar to the one against Indexical → for Rational Agents would damage it – the more far-reaching the result, the more damaging it would be. I suspect that a powerful result is there for the taking, since like PCCP, DAB equates a quantity involving conditional probability (V(A), expanded by the additivity rule) with an unconditional probability; and conditional probabilities are finer-grained quantities than unconditional probabilities for finite-ranged probability functions, as the original 'wallflower' result showed.

Oddie (1994) shows that any agent who (like Frederic) only distinguishes two levels of value, and who spreads credences evenly over a finite number of point-propositions, will assign values that find no match among the unconditional probabilities. There will be no ^{o} for such an agent. This is a good start, but again I would call such an agent a pathological case. Moreover, a more worthy anti-Humean target recognizes multiple levels of goodness, à la GDAB. A more powerful 'wallflower' result may be required to scotch that. After all, the possible values of the right-hand side are not restricted merely to the range of P – any cracks in its values can be filled to some extent by taking certain linear combinations of them, as per GDAB. But the problem may remain nonetheless: GDAB still equates a *conditional* expectation (of the utilities associated with various ways of realizing A, conditional on A) with an *unconditional* expectation (of various levels of goodness associated with A), and these beasts behave differently mathematically.

My result against the Indexical → versions exploited the fact that for finite-ranged probability functions, conditional probabilities (being ratios of unconditional probabilities) are more finely grained than unconditional probabilities, and the former can fall into cracks between the latter. If the probability function is *full*, there are no cracks – a fact that van Fraassen exploited in his positive result for Stalnaker's Thesis. He even showed that the → that conforms to the thesis for such functions can sustain a logic CE that may make it worthy of being called a conditional. For all that we have seen, a similar positive result for Desire-as-Belief may yet be in the offing. Perhaps sufficiently fine-grained utility and probability functions can sustain not only a ^{o}, but one with a rich enough structure to make it worthy of being called a 'goodness' operator. If so, Lewis could not so easily disparage "inconstancy" as merely a "little trick."

No-truth-value accounts of conditionals have analogues in no-truth-value accounts of goodness. For example, ^{o} could be understood along expressivist lines. In that case, $P(\mathring{A})$ could be interpreted not as the probability of $\mathring{A}$'s truth, but rather as a measure of a probability-like attitude to $\mathring{A}$ (recall

Adams's treatment of 'probabilities' of conditionals). In that case, the triviality results against Desire-as-Belief will not have any purchase.

27.4.3 A Final Twist

Lewis considered some interesting alternatives to DAB, but he was convinced that they were not viable anti-Humean positions. The first move was one of fine-graining:

(Fine-grained DAB) $V(A) = \Sigma_i \, g_i \, P(\mathring{A}_j)$

He argued that this would meet the same demise as DAB, and I have supported that with my perturbation argument. Another move was taking conditional probabilities:

(Desire as Conditional Belief) $V(A) = P(\mathring{A} \mid A)$

He argued that this was equivalent to Desire by Necessity, which he apparently found unacceptable. But what happens if we combine the two moves – fine-graining *and* taking conditional probabilities? We get:

(Fine-grained Desire as Conditional Belief) $V(A) = \Sigma_i \, g_i \, P(\mathring{A}_i \mid A)$

Now we have a happy alignment of conditional expectations on both sides of the equation (recall that $V(A)$ is a conditional expectation).

The $\{\mathring{A}_i\}$ form a partition, which we could just as well symbolize $\{Z_i\}$. The g_i represent levels of goodness. Make the identifications:

(Value Identifications) $V(A\mathring{A}_i) = g_i$

After all, $\mathring{A}_i$ is the proposition that A would be good to degree g_i. Then substituting into Fine-grained Desire as Conditional Belief, we have

$V(A) = \Sigma_i \, V(AZ_i) \, P(Z_i \mid A)$

Does this remind you of anything? It is none other than the additivity formula for evidential decision theory! Now go through the steps in reverse. Starting with the additivity formula, taking our partition to be $\{\mathring{A}_j\}$, and making the Value Identifications, we derive (Fine-grained Desire as Conditional Belief).

Far from this anti-Humean position being incompatible with evidential decision theory, given our identification of the g_i, the anti-Humean position is *entailed by* evidential decision theory! Ironically, while Lewis used evidential decision theory to refute an anti-Humean position, we have used it to prove another, closely related one. I wonder what Lewis would have made of this.

Acknowledgments

Thanks especially to Rachael Briggs, David Etlin, Leon Leonteyev, Daniel Nolan, Paolo Santorio, Wolfgang Schwarz, and Clas Weber, for very helpful discussion and comments – and to David Lewis, *sine qua non*.

Notes

1 Lewis (1986, 1988) formulates Jeffrey conditioning in non-standard ways, but they are equivalent to the more standard formulations that I have given.
2 Here, 'assertable' should be understood epistemically – 'acceptable' or 'assentable' might convey this better – rather than as a matter of the appropriateness of uttering the sentence. That depends on all sorts of pragmatic factors (social norms, Gricean norms, etc.) that are irrelevant here.
3 Lewis (1976) in fact goes on to give an alternative explanation of Adams's Thesis, based on Gricean norms of assertability.
4 This assumption may be problematic, since V need not be bounded.
5 I thank Daniel Nolan for suggesting this and the next reason for the importance of Desire-as-Belief to Lewis.
6 Kotzen (n.d.) makes this point.
7 I am grateful to David Etlin for pointing this out to me.

References

Adams, E. (1975). *The logic of conditionals*. Dordrecht: Reidel.
Appiah, A. (1985). *Assertion and conditionals*. Cambridge: Cambridge University Press.
Bennett, J. (2003). *Conditionals*. Oxford: Oxford University Press.
Bradley, R. & List, C. (2009). Desire-as-belief revisited. *Analysis*, 69(1), 31–7.
Byrne, A. & Hájek, A. (1997). David Hume and decision theory. *Mind*, 106, 411–28.
Edgington, D. (1995). On conditionals. *Mind*, 104, 235–329.
Etlin, D. (2009). The problem of noncounterfactual conditionals. *Philosophy of Science*, 76(5), 676–88.
Gibbard, A. & Harper, W. (1978). Counterfactuals and two kinds of expected utility. In C. Hooker, J. Leach & E. McClennen (Eds.), *Foundations and applications of decision theory*, pp. 125–62. Dordrecht: Reidel.
Hájek, A. (1989). Probabilities of conditionals – revisited. *Journal of Philosophical Logic*, 18, 423–8.
Hájek, A. (1994). Triviality on the cheap? In E. Eells & B. Skyrms (Eds.), *Probability and conditionals*, pp. 113–40. Cambridge: Cambridge University Press.
Hájek, A. (2012). The fall of Adams' thesis? *Journal of Language, Logic, and Information*, 21(2), 145–61.
Hájek, A. & Hall, N. (1994). The hypothesis of the conditional construal of conditional probability. In E. Eells & B. Skyrms (Eds.), *Probability and conditionals*, pp. 75–111. Cambridge: Cambridge University Press.
Hájek, A. & Pettit, P. (2004). Desire beyond belief. *Australasian Journal of Philosophy*, 82, 77–92.
Hall, N. (1994) Back in the (CCCP). In E. Eells & B. Skyrms (Eds.), *Probability and conditionals*, pp. 141–60. Cambridge: Cambridge University Press.
Harper, W. (1981). A sketch of some recent developments in the theory of conditionals. In W. Harper, R. Stalnaker & G. Pearce (Eds.), *Ifs*, pp. 3–38. Dordrecht: Reidel.
Harper, W. & Hooker, C. (Eds.) (1976). *Foundations of probability theory, statistical inference and statistical theories of science*, Vol. I. Dordrecht: Reidel.
Jeffrey, R. (1983). *The logic of decision*, 2nd ed. Chicago: University of Chicago Press.
Kaufman, S. (2004). Conditioning against the grain. *Journal of Philosophical Logic*, 33(6), 583–606.
Kotzen, M. (n.d.) Lewis's argument against desire-as-belief. Unpublished manuscript.
Levi, I. (1967). Probability kinematics. *British Journal for the Philosophy of Science*, 18, 197–209.
Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.
Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *Philosophical Review*, 85, 297–315.
Lewis, D. (1981) Causal decision theory. *Australasian Journal of Philosophy*, 59, 5–30.
Lewis, D. (1986). Probabilities of conditionals and conditional probabilities II. *Philosophical Review*, 95, 581–9.
Lewis, D. (1988). Desire as belief. *Mind*, 97, 323–32.
Lewis, D. (1989). Dispositional theories of value. *Proceedings of the Aristotelian Society*, 63, 113–37.
Lewis, D. (1996). Desire as belief II. *Mind*, 105, 303–13.

MacFarlane, J. (2005). The assessment sensitivity of knowledge attributions. In T. Szabo Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology*, Vol. 1, pp. 197–233. Oxford: Clarendon Press.
McGee, V. (1985). A counterexample to modus ponens. *Journal of Philosophy*, 82, 462–70.
McGee, V. (1989). Conditional probabilities and compounds of conditionals. *Philosophical Review*, 98, 485–541.
Oddie, G. (1994). Harmony, purity, truth. *Mind*, 103, 451–72.
Price, H. (1989). Desire as conditional belief. *Mind*, 98, 119–27.
Ramsey, F. (1965). *The foundations of mathematics (and other logical essays)*. London: Routledge and Kegan Paul.
Stalnaker, R. (1970). Probability and conditionals. *Philosophy of Science*, 37, 64–80.
Stalnaker, R. (1976). Letter to van Fraassen. In Harper & Hooker (1976), pp. 302–6.
van Fraassen, B. (1976) Probabilities of conditionals. In Harper & Hooker (1976), pp. 261–301.

28

Decision Theory after Lewis

JOHN COLLINS

Decision theory is no more than a systematic account of an ordinary, common sense philosophy of mind. It is a tool to be applied, for example, to the task of coming to know ourselves and others as persons, the task David Lewis describes in the 1974 paper of that name as "radical interpretation," where he envisages us "starting from scratch" and, given merely the facts about an agent as a physical system, "solv[ing] for the rest," that is, for the facts about what the agent believes, desires, and means. In particular, decision theory is of help in solving for the first two via a principle of rationalization: the beliefs and desires we ascribe to an agent ought to constitute good reasons for her observed behavior, and as a precondition for that: "those ascribed beliefs and desires should be coherent enough to permit the comparison of expected utilities of alternative ways of behaving" (Lewis 1974, 337).

The appeal to decision theory at this point in the interpretation process does not violate the requirement that the constraining conditions should only be drawn from an ordinary, common sense theory of persons, for, as Lewis continues:

> Decision theory (at least if we omit the frills) is not an esoteric science, however unfamiliar it may seem to an outsider. Rather it is a systematic exposition of the consequences of certain well-chosen platitudes about belief, desire, preference and choice. It is the very core of our common-sense theory of persons, dissected out and elegantly systematised.

In the paper quoted, Lewis cites the version of the formal theory developed in Richard Jeffrey's (1965) *The Logic of Decision* (the first edition of Jeffrey 1983). Lewis's own most systematic published exposition of decision theory comes in the 1981 piece "Causal Decision Theory."

28.1 Two Versions of Decision Theory

Given a set W of possible worlds, we assume that a potentially rational agent has, at each moment, a credence function C and a value function V. The *credence* function is simply a probability function

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

over the algebra of subsets of W. Each subset X of W may be thought of as a *proposition*, namely the proposition that is true at each world $w \in X$. The credence $C(X)$ is a measure of the agent's degree of belief in the proposition X.

The value function V is defined initially over individual possible worlds. $V(w)$ is a measure of "how satisfactory it seems to the agent for w to be the actual world" (Lewis 1981, 6).

Our aim is to extend the value function from individual worlds to propositions so that it gives a measure of degrees of desire. Decision theorists (all of them) agree that desirability, that is, subjective value, is to be explicated as expected value. They agree that insofar as decision theory is a normative enterprise, it prescribes that one choose so as to maximize expected value. Furthermore, they agree that the notion of expected value involves, at least implicitly, the notion of revision of degrees of belief. The expected value of an option A is the credence-weighted average of the ways in which A might be true. The credences involved are just those that result when the agent revises to accommodate the proposition that A is the choice made. In other words, to calculate the expected value of A, the agent should revise to accommodate A, and see how good things would be then.

$$\text{Expected Value: } EV(A) = \sum_{w \in A} C_A(w).V(w)$$

The sum here is over all the A-worlds and C_A is the function that results when the agent's credences C are adjusted so as to accept A. To go easy on the reader's eye I am abusing notation by not distinguishing worlds from point propositions, and for simplicity I am assuming here that there are only finitely many possible worlds.

Note that this definition involves a revised credence function C_A: C revised to accommodate the truth of A. This is what I meant above when I said that the notion of expected value involves at least implicitly the notion of revision of belief. The dispute between the two versions of decision theory amounts to no more than a disagreement about what sort of revision is appropriate here. There are two rival suggestions about revision method, hence two candidates for the role of expected value, and hence two versions of decision theory. If revision in the above neutral definition is taken to be revision by conditionalization, then we get evidential decision theory:

$$\text{Evidential Expected Value: } EEV(A) = \sum_{w \in A} C(w/A).V(w)$$

If the revision in question goes by an alternative method known as "imaging," then we get what is usually called the "causal theory":

$$\text{Causal Expected Value: } CEV(A) = \sum_{w \in A} C^{\#}_A(w).V(w)$$

Conditional credence is defined as a quotient of credences:

$$C(X/Y) = C(X\&Y)/C(Y)$$

provided that $C(Y) > 0$ so that the quotient is well-defined. Conditionalization, the revision process that replaces C with $C(-/Y)$, is the familiar Bayesian method of updating credence when one comes to fully accept the truth of the proposition Y. A useful way of visualizing the method is to conceive

of conditionalization on Y as a matter of simply "wiping away" all of the probability that C assigns to worlds at which Y is false, and then renormalizing the remaining probability so that it sums to 1. This renormalization process leaves fixed the ratios of credences that the agent assigns to propositions that entail Y.

Imaging on A is a less familiar process. It may be thought of this way. Consider each $\neg A$-world, that is, each of the worlds at which A is false, to which C assigns non-zero credence. For each such world v find the "closest" or "most similar" A-world to v, call it "$v\#A$" and shift the credence C assigned to v over to $v\#A$.

The sense of "closest" or "most similar" here is intended to be that familiar from the Lewis–Stalnaker semantics for the subjunctive conditional. If one follows Lewis in allowing that there might be many A-worlds in a tie for closest A-world to v, then the way to give a formal development of imaging is as follows.

Say that an imaging method assigns to each world v and proposition A a credence function $v^{\#}{}_{A}$ such that $v^{\#}{}_{A}(A) = 1$, that is, such that $v^{\#}{}_{A}$ assigns positive credence only to A-worlds (intuitively: only to the A-worlds that are closest to v).

The imaging method may then be extended from worlds to credence functions C via the definition:

$$C^{\#}_{A}(w) = \sum_{v} C(v).v^{\#}_{A}(w)$$

The credence function $C^{\#}{}_{A}$ so defined is said to be the result of imaging C to accept A. It is also referred to as the A-image of C. Note that this more general development includes the earlier kind of imaging as the special case in which for each v, the image $v^{\#}{}_{A}$ is an *opinionated* credence function that assigns probability 1 to a single world. Call these *sharp* imaging methods.

We now have two competing accounts of expected value, based on two belief revision methods. But what, exactly, is at stake in the dispute between these competing accounts?

The debate between the proponents of these so-called "evidential" and "causal" versions of decision theory usually focuses on the following highly unusual kind of choice situation:

Newcomb's Problem

> There are two boxes on the table in front of you. One of them is transparent and can be seen to contain one thousand dollars. The other is opaque. You know that it contains either one million dollars or nothing. You must decide whether to take (1) only the contents of the opaque box (call this the *One-Box option*); or, (2) the contents of both boxes (the *Two-Box option*). You know that a remarkably accurate Predictor of human deliberation placed the million dollars in the opaque box yesterday if and only if she then predicted that you would choose today to take only the contents of that box. You have 90 percent confidence in the Predictor's reliability. What should you do? (See Nozick 1969.)

We may represent the problem with the following pair of matrices. In each matrix the rows correspond to your options, and the columns to states of the world you take to be relevant to your choice problem. "M" stands for the proposition that the Predictor placed the million dollars in the opaque box yesterday, or, equivalently, that she then predicted you would make the One-Box choice. A and B are the One- and Two-Box options respectively. The first matrix gives your (the agent's) credences, where x is whatever degree of belief you initially assign to the proposition that you will end up making the One-Box choice. Your confidence in the accuracy of the Predictor is reflected by the fact that you have conditional credences: $C(M/A) = C(\neg M/B) = 0.9$. The second is the value matrix; we shall

assume that your values for outcomes are given by the dollar amounts received in each of the four possible cases.

	M	$\neg M$
A	$0.9x$	$0.1x$
B	$0.1(1-x)$	$0.9(1-x)$

Credences ($0 \leq x \leq 1$)

	M	$\neg M$
A	1,000,000	0
B	1,001,000	1,000

Values

Then the Evidential Expected Values are calculated as follows:

$$\mathrm{EEV}(A) = C(M/A).V(M\&A) + C(\neg M/A).V(\neg M\&A) = 900,000$$

$$\mathrm{EEV}(B) = C(M/B).V(M\&B) + C(\neg M/B).V(\neg M\&B) = 10,000$$

and so according to the evidential theory the One-Box choice A maximizes expected value.

But this prescription seemed to Lewis, as it seems to many of us, wrong. There is a straightforward dominance argument supporting the Two-Box choice. The prediction was made yesterday. The million dollars is either there in the opaque box or it is not. Nothing that you do now can change the situation. Whichever situation now obtains, you will be one thousand dollars better off if you take both boxes. You should therefore choose to take both boxes.

Of course dominance reasoning is valid only when outcomes are independent of choice. But whether or not this counts against such reasoning here depends upon what one means by "dependence." What makes the Newcomb Problem interesting is that it is one in which outcomes are probabilistically dependent on, yet causally independent of, the agent's options. Two-Boxers like Lewis think that causal independence is sufficient to underwrite the dominance argument.

And of course the Two-Boxer needs to show where the evidential theorist's expected value argument has gone wrong. That is precisely the point of the version of expected value which appeals to imaging rather than to conditionalization as the implicit belief revision method.

Recall that the idea of imaging on a proposition A was to transfer credence away from each $\neg A$-world and reassign it to the closest A-world. The sense of "closest" or "most similar" here is intended to be that familiar from the Lewis– Stalnaker semantics for the subjunctive conditional. Thought of this way, it is not hard to see that $C^{\#}{}_{A}(w)$ is none other than the prior probability $C(A\Box\!\!\rightarrow w)$ assigned by the agent to the subjunctive conditional $A\Box\!\!\rightarrow w$. Since such conditionals are widely regarded to be relevant to a proper analysis of the causal relation, this definition came to be thought of as yielding a notion of expected value sensitive to the causal structure of the situation in question. Hence the name "causal decision theory." In fact, however, I think that that is the wrong way to look at things. I'll describe what I take to be the right way later.

In the meantime let's see how a notion of expected value based on imaging can underwrite the Two-Box choice in Newcomb's Problem. Recall the earlier matrix representation:

	M	$\neg M$
A	$0.9x$	$0.1x$
B	$0.1(1-x)$	$0.9(1-x)$

Credences ($0 \leq x \leq 1$)

	M	$\neg M$
A	1,000,000	0
B	1,001,000	1,000

Values

When one images on each of the agent's options in turn, the facts about whether or not the million dollars is in the opaque box stay fixed. That is because that fact is causally independent of what you choose to do. So, for example, in the world of the top left cell in which the million dollars is there and you make the One-Box choice, the following subjunctive conditional is true: If you had made the Two-Box choice, the million dollars would still have been there. In other words the world $M\&A$ images on B to the world $M\&B$. By similar reasoning $\neg M\&A$ images on B to $\neg M\&B$ and we can proceed to obtain the following matrices for the credence functions revised by imaging on each of the agent's two options in turn:

	M	$\neg M$
A	$0.8x + 0.1$	$0.9 - 0.8x$
B	0	0

Credences after imaging on A

	M	$\neg M$
A	0	0
B	$0.8x + 0.1$	$0.9 - 0.8x$

Credences after imaging on B

Then the causal expected values are calculated as follows:

$$\mathrm{CEV}(A) = C^{\#}_A(M\&A).V(M\&A) + C^{\#}_A(\neg M\&A).V(\neg M\&A) = 800,000x + 100,000$$

$$\begin{aligned}\mathrm{CEV}(B) &= C^{\#}_B(M\&B).V(M\&B) + C^{\#}_B(\neg M\&B).V(\neg M\&B) = 800,800x + 100,100 + 900 \\ &\quad - 800x = 800,000x + 101,000\end{aligned}$$

and so, in accordance with the dominance argument, the Two-Box choice is seen to be one thousand dollars more attractive than the One-Box choice when expected value is calculated this way.

28.2 Imaging and Dependency Hypotheses

Now in fact the causal decision theory outlined in the previous section is closer to the way the theory was presented by J. Howard Sobel (1978, 1994) than to the official version that Lewis gives in Lewis (1981). Lewis preferred a development that appealed to what he called *dependency hypotheses* but thought of his own account and Sobel's as essentially equivalent. Let's see how this equivalence works. (Note to the reader: this section, which is more technical than the rest of the piece, may safely be skipped over without losing track of the main narrative.)

A *dependency hypothesis* K is a "maximally specific proposition about how the things [the agent] cares about do and do not depend causally on his present actions" (Lewis 1981, 11). The possible dependency hypotheses form a partition of the set of possible worlds because of their maximal specificity and because they cannot differ without conflicting. Exactly one such hypothesis is true at each world. Lewis thought that if you were certain of which dependency hypothesis holds in the actual world, then you could simply follow the dictates of evidential decision theory. However, if your credence is spread over several hypotheses, "you should consider the expected value of your options under the several hypotheses; you should weight these by the credences you attach to the hypotheses; and you should maximize the weighted average" (1981, 11).

This yields the following definition of causal expected value, or, as Lewis preferred to call it in this context: *expected utility* U.

$$U(A) = \sum_K C(K).V(A\&K)$$

Here the sum is over all of the dependency hypotheses K, and the $V(A\&K)$ is simply the evidential expected value of the proposition $A\&K$. Note that, in order for this latter value to be well-defined, it must be the case that $C(A\&K) > 0$ for each K.

How do we get from this to the formulation in terms of imaging?

First of all, expand the expression $V(A\&K)$ as a conditional credence weighted average of values of individual worlds. Thus:

$$V(A\&K) = \sum_w C(w/A\&K).V(w)$$

then, plugging that into the definition of $U(A)$ and rearranging the order of summation, we obtain:

$$U(A) = \sum_w \sum_K C(K).C(w/A\&K).V(w)$$

We can now define a world-by-world imaging method. Let the world v image on A to the credence function:

$$v^{\#}_A(w) =_{\text{def}} C(w/A\&K_v)$$

where K_v is the dependency hypothesis true at world v. And this definition extends, as we saw earlier, to an imaging method on credence functions, since:

$$C^{\#}_A(w) = \sum_v C(v).v^{\#}_A(w) = \sum_v C(v).C(w/A\&K_v) = \sum_K C(K).C(w/A\&K)$$

Thus it can be seen that for this way of defining an imaging method # in terms of dependency hypotheses, the expression for $U(A)$ above that involves the double sum reduces to:

$$U(A) = \sum_w C^{\#}_A(w).V(w)$$

which was precisely our earlier definition of causal expected value.

Say that two worlds v and w *image alike* on the agent's options when for all A: $v^{\#}{}_A = w^{\#}{}_A$. Then Lewis's dependency hypotheses are revealed to be equivalence classes of worlds under the relation of imaging alike.

Or at least this is true when the imaging method is well-behaved. One has to be careful, as is illustrated by the following example:

Example: Suppose we have four worlds u, v, w, and x. Let $A = \{u, v\}$, $B = \{w, x\}$ be the agent's options. Suppose that w is the closest B-world to u, but that v rather than u is the closest A-world to w. Let x be the closest B-world to v and v the closest A-world to x. Define an imaging method # in terms of these relations of comparative closeness. Then the equivalence classes of worlds under the relation of imaging alike on the agent's two options are: $\{u\}$, $\{w\}$, and $\{v, x\}$. But now if we take these three equivalence classes to be our dependency hypotheses and apply Lewis's definition of expected utility to the example we get:

$$U(A) = \sum_K C(K).V(A\&K) = C(u).V(u) + C(w).V(A\&w) + C(\{v, x\}).V(v)$$

and that expression is ill-formed since $A\&w = \emptyset$ and so $V(A\&w)$ is undefined.

In order to construe dependency hypotheses as equivalence classes of worlds that image alike, we have to restrict our attention to imaging methods that are well-behaved in the following sense:

> Definition: Say that an imaging method is *well-behaved* when for all worlds w and all options A, $w^{\#}{}_{\Lambda}(K_w) = 1$, where K_w is the equivalence class of w under the relation of imaging alike on the agent's options, that is, when the credence initially assigned to w is completely relocated to worlds within the dependency hypothesis true at w.

28.3 Desire as Belief

The second of the two disputes on which this chapter centers concerns the possibility of giving a coherent cognitivist or anti-Humean account of motivation. Our ordinary explanations of behavior make reference to two kinds of mental state, which we call belief and desire. It is worth asking whether these kinds are necessarily distinct, or whether it might be possible to construe desire as belief of a certain sort – belief, say, about what would be good. Decision theory formalizes our notions of belief and desire, treating each as a matter of degree. In this context the thesis that desire is belief might amount to the claim that the degree to which an agent desires any proposition equals the degree to which the agent believes that it would be good if that proposition were true.

In Lewis (1988) David Lewis presented an argument against this anti-Humean proposal that desire is belief. Lewis proved that, on pain of triviality, the Desire-as-Belief Thesis cannot be added to the axioms of decision theory.

If the Desire-as-Belief Thesis were true, it would be possible to do away with all reference to desire in ordinary explanations of behavior. Talk of desire could systematically give way to talk about belief in objective goodness. Two aspects of this are worth noting:

1. The proposal would reduce an apparently non-cognitive attitude (desire) to one that is clearly cognitive (belief).
2. The cognitive attitude offered is an attitude towards something objective (goodness), whereas the original attitude (the desire) was something subjective.

Seen in this light, the Desire-as-Belief Thesis appears as but one of a whole slew of similar proposals sharing with it one or both of those features. Here are some examples:

(a) Credence as belief about objective chance.
(b) Epistemic possibility as belief about objective possibility.
(c) Conditional belief as belief about objective conditionals.
(d) Desire as belief about objective goodness.
(e) Hope as belief about objective hopefulness.
(f) Terror as belief about objective ghastliness.

And so on. Readers may extend this list for themselves *ad libitum*. Some of these proposals are pretty clear non-starters. Take, for example, (a), which might be formulated as follows, using *Ch* for objective chance.

Credence-as-Belief: $C(X) = x$ iff $C\,(Ch(X) = x) = 1$

This thesis falls to the "Integrating Out Argument." Suppose that the agent's credence is divided between various chance hypotheses. For simplicity, let's assume it to be divided evenly between just two hypotheses: the hypothesis that the chance of X is 0.9, and a contrary hypothesis that the chance of X is 0.1. But then:

$$\begin{aligned} C(X) &= 0.5 \times C\,(X/Ch(X) = 0.9) + 0.5 \times C\,(X/Ch(X) = 0.1) \\ &= (0.5 \times 0.9) + (0.5 \times 0.1) \\ &= 0.5 \end{aligned}$$

where we have appealed to the plausible principle that:

$$C\,(X/Ch(X) = x) = x$$

By the Credence-as-Belief Thesis, this entails that $C\,(Ch(A) = 0.5) = 1$, which contradicts our initial assumption that the agent's subjective probability is divided between two distinct chance hypotheses. The Credence-as-Belief Thesis cannot be maintained, as it implies complete opinionation about chance. Similar reasoning prevails against thesis (b).

Now in fact the Integrating Out Argument can also be used against a simple formulation of the thesis that desire is belief, according to which the agent desires X to degree x just in case the agent fully believes that X's degree of objective goodness is x. We suppose that the function V gives the agent's degrees of desire, and that there is a function g that assigns to each proposition X its degree of objective goodness. As before, the agent is taken to have full belief in X iff X receives subjective probability 1.

Desire-as-Full-Belief: $V(X) = x$ iff $C(g(X) = x)) = 1$

The *reductio* proceeds, as before, from the assumption that the agent entertains two rival hypotheses about the goodness of some proposition. We see that this first version of the DAB Thesis entails that any rational agent must be completely opinionated about matters of objective goodness. That is an absurd conclusion.

But the cognitivist may object that this first version of the Desire-as-Belief Thesis, and corresponding first versions of our other theses, are misformulations. Belief, as well as goodness, admits of degree. Two dimensions of degree allow a three-way ambiguity in sentences like:

> To what extent do you believe that a tightening of the money supply would adversely affect the stock market?

On the first interpretation, the question takes for granted full belief in the proposition that the effect of monetary tightening on the markets would be adverse, and is asking you to judge the extent of the damage. On the second reading, the question is asking you to judge your degree of belief in the proposition that stock prices would fall if the Fed cut back reserves. On the third reading, the word "extent" serves to invoke simultaneously consideration of degree of belief and of the magnitudes of the possible shifts in the market index. In this third case, an appropriate answer to the question might involve something like an expectation: a credence-weighted average of the magnitudes of the changes deemed possible.

There is a corresponding range of possible formulations of each of the above theses. Our first version of the Desire-as-Belief Thesis, in which only goodness is taken to be a matter of degree, corresponds to the first reading of the question. Corresponding to the second interpretation of the question is a second version of DAB, in which belief rather than objective goodness is allowed to admit of degree.

$V(X) = x$ iff $C(g(X) = 1) = x$

Since we are now taking objective goodness to be an all or nothing matter, we may suppose for the purposes of reformulating the second version of the DAB Thesis that to each proposition X there corresponds a proposition X° (pronounced "X-halo"): the proposition that X is objectively good. The second version of the DAB Thesis simplifies to:

Desire-as-(Degree of) Belief: $V(X) = C(X^\circ)$

The third version of the DAB Thesis allows that both belief and objective goodness come in degrees. Degree of desire is then equated with expectation of goodness; it is represented as a credence-weighted average of all the possible degrees of objective goodness. Call this the:

Desire-as-Expectation Thesis: $V(X) = \Sigma_x x.C(g(X) = x)$

The DAE Thesis appears to be the most general of the three versions we have considered, but it turns out our second version is already general enough. By this I mean that there is really nothing to be gained by moving from the second version to the third. We will follow Lewis in taking the second, Desire-as-Degree-of-Belief, version of the thesis to be our official statement of DAB. There will be no loss of generality since the discussion could be straightforwardly (though tediously) adapted to cover the Desire-as-Expectation formulation instead.

In the 1988 paper Lewis presented an argument against this anti-Humean proposal. Lewis proved that the DAB thesis cannot be added to the axioms of decision theory without trivializing the theory. The 1996 sequel "Desire as Belief II" presents a simpler, more streamlined version of the argument. See also the paper by Arló Costa, Collins and Levi (1995). The argument is closely related to another famous trivialization result that Lewis proves in (1976).

The first step is to see that DAB is equivalent to the conjunction of the following two principles:

Desire-as-Conditional-Belief: $V(X) = C(X^\circ/X)$

Independence: $C(X^\circ/X) = C(X^\circ)$

Proof: If DAB is assumed to hold throughout a set of credence-value function pairs (C, V) closed under conditionalization, we may conditionalize by X on both sides of the equation. This yields DACB. IND then follows directly from DACB and DAB. The implication in the other direction is immediate.

Desire-as-Belief Theorem: Desire-as-belief implies near-opinionation or indifference.

Proof: An opinionated agent is one who assigns credence 1 to a single possible world. Say that an agent is "near-opinionated" if her credence function assigns positive credence to no more than two worlds. An indifferent agent is one whose value function V is constant over all possible worlds. If an

agent is neither indifferent nor near-opinionated, we may choose a proposition X such that the agent's credence function C assigns positive credence to all three of the propositions: $X\&\neg X^{\circ}$, $X\&X^{\circ}$, and $\neg X\&X^{\circ}$. Now if IND holds for C, it will fail after conditionalization on $X \vee X^{\circ}$, since this updating raises the probability of X, but not the conditional probability $C(X^{\circ}/X)$. QED.

This is a surprising result. How should one respond? What we might call the "straight response" would be to take the DAB Theorem as establishing that the simplest version of a rationalist, anti-Humean account of motivation is actually incoherent.

But one may reasonably doubt whether that is something that could really be demonstrated so easily by a little formal argument of the sort just given. Christine Korsgaard (1986) has drawn a distinction between two kinds of skepticism about the rationalist position. A weaker non-cognitivist view, which Korsgaard calls content skepticism, finds fault with particular, specific principles that have been advanced as candidate norms of practical reason. A more ambitious Humean strategy aims at establishing what Korsgaard calls motivational skepticism. According to motivational skepticism, we can see *a priori* that there can be no such thing as pure practical reason, even before coming to consider any specific rationalist proposals. The straight response would be to take the Desire-as-Belief Theorem as an *a priori* argument for the Humean theory, as an argument for the Humean position based only on a consideration of the logic of desire and belief, or, in other words, as a conclusive argument for motivational skepticism.

Korsgaard rejects motivational skepticism, as, presumably, do all those who propose specific norms for the operation of pure practical reason. I share their doubts about the possibility of this kind of strong formal proof of the Humean theory of motivation. But that is just to say that I believe we have reason to resist what I have been calling the "straight response."

I shall proceed now to investigate what it would take to deny the conclusion of the Desire-as-Belief Theorem. In other words, let us assume that cognitivism is not a position that can be ruled out on purely decision theoretic grounds. Let's hold the Desire-as-Belief Thesis fixed for the moment and see if we can reconstrue. Lewis's argument as a reduction against one of that argument's premises or presuppositions.

Call a credence revision method "#" *linear* if it satisfies the following condition:

Linearity: If $C = \alpha.C' + (1-\alpha).C''$, then $C^{\#}_{A} = \alpha.C'^{\#}_{A} + (1-\alpha).C''^{\#}_{A}$.

According to this linearity condition the operations of revising a credence function and taking a mixture of credence functions commute. In other words, the linearity condition is saying that the revision of a mixture of two credence functions is the mixture of the revisions.

The demand for linearity is the demand that the action of the credence revision method be completely determined by the action of that method on each point proposition to which the function to be revised assigns non-zero credence. When a revision method has this property of linearity, that is, of having its action completely determined by its action on point propositions, it is harmless and helpful to abuse notation a little further by conflating each world w with the opinionated credence function that assigns probability 1 to world w and takes the value 0 elsewhere. I shall further simplify matters by restricting our attention to what I earlier called *sharp* revision methods. A revision method is said to be sharp if it "preserves opinionation," that is, if it maps any opinionated credence function to another opinionated function.

When the revision method is sharp, the Linearity condition may be equivalently expressed in the following way:

$$C^{\#}_{A}(X) = C(\{w : w\#A \in X\})$$

All of this simplification and notation abuse is intended just to enhance the readability of the following important result:

Linearity Theorem: If an agent's degrees of desire are given by expected values calculated according to some sharp, linear revision method, then they satisfy the Desire-as-Belief Thesis.

Proof: Suppose the agent's credence function is C, and that the agent's value measure is V that assigns to each world either the value 0 or 1. Let $G = \{w : V(w) = 1\}$. Let $A^\circ = \{w : w\#A \in G\}$. Then:

$$\begin{aligned} V(A) &= \sum_{w \in A} C_A^\#(w).V(w) \\ &= \sum_{w \in G} C_A^\#(w) \\ &= C_A^\#(G) \\ &= C(\{w : w\#A \in G\}) \\ &= C(A^\circ) \end{aligned}$$

This is an interesting result indeed! It tells us that the imposition of the linearity condition on the revision method implicit in the definition of expected value is sufficient to achieve the anti-Humean aim of coherently construing desire as a cognitive attitude toward an objective feature of the world.

How does this fit in with the earlier proof that Desire-as-Belief implies opinionation or indifference? Remember that that proof assumed that degree of desire or expected value was defined the way the evidential decision theorist says it should be, in other words with the agent's probabilities revised by conditionalization. Since conditionalization is clearly not a linear revision method, there is no conflict between the earlier result and what we have just proved. Still, it would be a useful exercise to go back and check exactly where Lewis's proof of the Desire-as-Belief result breaks down (as it must!) when the revision method is linear.

The answer to that exercise is that, if the revision method is linear, the argument breaks down at the very first step, where Desire-as-Conditional-Belief was derived from Desire-as-Belief. That step appealed to an Invariance Assumption, that is, to the assumption that $V_A(B) = V(A\&B)$, to obtain the result that $V_A(A) = V(A)$. Now the Invariance Assumption holds for revision by conditionalization, but not for doxastic changes in value that are the result of linear revisions of the agent's degrees of belief.

It is now high time to reveal what was perhaps already obvious from the choice of notation: linearity is a characteristic of imaging.

Theorem: Imaging methods are linear.
Proof: Suppose that $C = \alpha.C' + (1 - \alpha).C''$. Then

$$\begin{aligned} C_A^\#(w) &= \sum_v C(v).v_A^\#(w) \\ &= \sum_v (\alpha.C'(v) + (1-a).C''(v)).v_A^\#(w) \\ &= a\sum_v C'(v).v_A^\#(w) + (1-a)\sum_v C''(v).v_A^\#(w) \\ &= aC_A'^\#(w) + (1-a)C_A''^\#(w) \end{aligned}$$

In light of this result, it would appear that the so-called "Causal Decision Theory" might more appropriately be named Linear Decision Theory.

The debate between the evidential and causal theories of decision usually focuses on wacky, science-fiction-like scenarios like the Newcomb's Problem example discussed earlier, But judgments about such cases differ, and while the story in question strikes some as a reliable intuition pump, to others (and it's hard to improve here on Dick Jeffrey's memorable phrase!) it seems to be no more than a "Prisoner's Dilemma for space cadets."

The moral I draw from the discussion of this section is that perhaps the debate over Newcomb-like problems can simply be side-stepped. It may be that the strongest argument in favor of a Causal (or Linear) Decision Theory is that it is the only kind of decision theory that leaves room for a kind of cognitivism about value that surely should not be refutable via a simple formal argument about the logic of belief and desire.

28.4 A Puzzling Feature of Lewis's Views

It may at this point strike you as just a little strange that David Lewis, author of the original argument against the possibility of Desire-as-Belief, and champion of all things Humean, should be the very same David Lewis who defends the only kind of decision theory which leaves open the possibility of cognitivism about motivation, a version of decision theory that deserves to be called "anti-Humean" at least as much as it deserves to be called "causal". I certainly find it puzzling.

Lewis was clearly not unaware of this issue, but I am not sure why he remained so unperturbed. The sort of line I have been pushing here is dismissed rather quickly in a parenthetical comment. Lewis says:

> (A famous difficulty need not concern us here. Suppose a certain action would serve as an effective means to your ends, yet at the same time it would constitute evidence – evidence available to you in no other way – that you are predestined inescapably to some dreadful misfortune. Should you perform the action? – Yes; your destiny is not a consideration, since that is outside your control. Do you desire to perform it? – No, you want good news, not bad. Since our topic here is not choiceworthiness but desire, and since the two diverge, we adopt an "evidential" conception of expected value, on which the value of the useful action that brings bad news is low. Choiceworthiness is governed by a different "causal" conception of expected value.) (1996, 304)

This strikes me as a rather weak line to take against the kind of cognitivist strategy I have outlined here. Having distinguished newsworthiness from choiceworthiness, is it really as clear as Lewis thinks that the ordinary folk-psychological notion of desire goes with the former rather than the latter? It is crucial to Lewis's interpretation of the formal result that this be the case, else the Desire-as-Belief Theorem so-called would really have nothing at all to tell us about desire. Is degree of desire really degree of newsworthiness, as Lewis so confidently claims? I was at first inclined to think quite the opposite. I would have thought that one's desires and beliefs were first and foremost revealed by one's dispositions to choose, and to act. I still think that desire is conceptually pretty tightly connected to the notion of choice. But I wouldn't want to insist on that here. In fact I suspect that it is simply an indeterminate matter whether the ordinary folk-psychological concept of desire lines up with the notion of newsworthiness or rather with choiceworthiness. The fact is that the newsworthy/choiceworthy distinction is not one that the folk are likely to make, and notoriously when it is forced upon the attention of the uninitiated, for example by asking them to contemplate some bizarre scenario like that in Newcomb's Problem, there is no univocal folk response. That a community of intelligent

folk who share a common folk psychological scheme can split so dramatically and irreconcilably on the subject of whether to choose one box or two boxes in the Newcomb's Problem suggests to me pretty strongly that the matter is one of conceptual indeterminacy. Similarly it seems to me that there are plenty of those who defend anti-Humean accounts of motivation, plenty of cognitivists, plenty of defenders of the notion of objective value and plenty of neo-Kantian rationalists on the topic of pure practical reason, who share, pretty much, a common folk-psychological scheme, and who are not simply guilty of some kind of conceptual confusion.

Recall the passage quoted earlier:

> Decision theory (at least if we omit the frills) is not an esoteric science, is a systematic exposition of the consequences of certain well-chosen platitudes about belief, desire, preference and choice. It is the very core of our common-sense theory of persons, dissected out and elegantly systematised.

I doubt whether it is really a "platitude," really a part of the "core of our common-sense theory of persons" that degree of desire is to be explicated by a notion of expected value that involves conditionalization rather than a linear probability revision method. Perhaps the whole evidential/causal debate is to count as "frills" rather than "core." But, if so, one can hardly adopt a firm stance on such an issue in the course of giving an argument that is supposed to be a strong *a priori* refutation of cognitivism.

In conclusion: I think that there is no small irony in the fact that the greatest contemporary defender of Humeanism is also one of the co-founders of the only version of decision theory that has the resources to make sense of the notion of objective value and of a cognitivist, anti-Humean account of desire.

References

Arló Costa, H., Collins, J. & Levi, I. (1995). Desire-as-Belief implies opinionation or indifference. *Analysis*, 55(1), 1–5.

Jeffrey, R. (1983). *The logic of decision*, 2nd ed. Chicago: University of Chicago Press.

Korsgaard, C. (1986). Skepticism about practical reason. *Journal of Philosophy*, 83(1), 5–25.

Lewis, D. (1974). Radical interpretation. *Synthese*, 23, 331–44. Reprinted in D. Lewis (1983), *Philosophical papers*, Vol. I, pp. 108–21. Oxford: Oxford University Press.

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *Philosophical Review*, 85, 297–315.

Lewis, D. (1981). Causal decision theory. *Australasian Journal of Philosophy*, 59(1), 5–30. Reprinted in D. Lewis (1986), *Philosophical papers*, Vol. II, pp. 305–37. Oxford: Oxford University Press.

Lewis, D. (1988). Desire as belief. *Mind*, 97, 323–32.

Lewis, D. (1996). Desire as belief II. *Mind*, 105, 303–13.

Nozick, R. (1969). Newcomb's problem and two principles of choice. In N. Rescher (Ed.), *Essays in honor of Carl G. Hempel* pp. 114–46. Dordrecht: Reidel.

Sobel, J.H. (1978). Chance, choice and action: Newcomb's problem resolved. Unpublished manuscript.

Sobel, J.H. (1994). *Taking chances: essays on rational choice*. Cambridge: Cambridge University Press.

29

Lewis on Mereology and Set Theory

JOHN P. BURGESS

David Lewis in the short monograph *Parts of Classes* (Lewis 1991, henceforth PC) undertakes a fundamental re-examination of the relationship between mereology, the general theory of parts, and set theory, the general theory of collections. He assumes a certain minimum background familiarity with both subjects, and limitations of space make the same assumption inevitable in the present account. Varzi (2011) is recommended as a clear, concise survey of mereology. Among the assumptions surveyed, Lewis's are about the strongest, including *unrestricted composition*, the claim that any things whatsoever have a *fusion*, but leaving open whether everything is a fusion of *atoms*, or things with no parts other than themselves. For set theory in the form in which it will be considered here, Boolos (1989, with Boolos 1971 in the background), to which reference will in any case be essential, should suffice.

Lewis became aware of the possibility of certain technical improvements too late in the production process for his book to incorporate them except as an appendix to PC (Burgess, Hazen and Lewis 1991, henceforth PC*), co-authored with A.P. Hazen (who has also provided many helpful comments on earlier drafts of the present study) and the present writer, and including also a contribution by W.V. Quine. Lewis became aware of the possibility of certain further technical improvements too late to incorporate them in the book at all, whence the follow-up paper, "Mathematics is Megethology" (Lewis 1993, henceforth MM), which Lewis described in its introductory session as "an abridgement of parts of [PC] not as it is, but as it would have been had I known sooner what I know now."

A half-dozen themes pursued in the book and article: (1) formulation of theses on how the mereological notion of part applies to classes; (2) restatement of traditional mereologists' complaints about set theory, concentrating on the notion of singleton sets, and motivating a "structuralist" approach to be further explained below; (3) defense beyond what is already found in Lewis (1986) of the background assumptions of mereology, including controversial theses on "ontological innocence" and "composition as identity"; (4) elaboration of a framework combining mereology with the plural quantification of Boolos (1984 and 1985); (5) consideration of how to simulate within such a framework, using assumptions related to the axiom of choice, quantification over relations; (6) development of the structuralist treatment of set theory using such simulated quantification, showing

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

that beyond the framework the only assumptions needed are about how many atoms there are, whence the identification of mathematics with "megethology," the theory of size.

I will take up the six themes in the order listed. The most important sources are as follows: PC chapter 1 for (1); PC chapter 2 for (2); PC §3.6 for (3); PC chapter 3 for (4); PC* plus Hazen (1997) for (5); and MM plus Boolos (1989) for (6).

29.1 Parts of Classes

The notion of "part" is applied in ordinary language to entities of many different sorts. As the introductory portion of Varzi (2011) makes plain, however, the word is not, in ordinary language, always applied in the same sense. To claim logic-like universality for mereology – such a claim is not fully explicit in Lewis, though his thought tends in that direction – is to claim there is a core sense of "part" that can be applied univocally to entities of any sort whatsoever, including classes, if such there be. It is the applicability of the notion of and the mereological theory of "parthood" to sets and classes that is Lewis's most fundamental claim.

The form of the theory of sets and classes of concern to Lewis is a versions of what is variously called "second-order Zermelo-Frankel set theory with choice" (henceforth second-order ZFC) or "Morse-Kelly set theory" (henceforth MK) to be met with in the literature. Both versions can be developed to as to admit individuals (German: *Urelemente*), items that are neither sets nor classes but can belong to such collections. In the version followed by Lewis, that of the *lumpers* as opposed to the *splitters*, sets are classes of a kind, the "small" ones and equivalently the ones that can be members of other classes; those of the other kind, the "large" ones and equivalently the ones that cannot be members of other classes, are called *proper classes*.

Lewis has one main thesis concerning mereology and classes, with one outstanding corollary. His main thesis follows immediately from two subordinate theses. The second of these is deduced from three yet further subordinate theses.

Main Thesis: *The parts of a class are precisely its subclasses.*
First Thesis: *One class is part of another if and only if it is a subclass.*
Second Thesis: *Any part of a class is a class.*
Division Thesis: *There are only individuals, classes, and fusions thereof.*
Priority Thesis: *No class is a part of any individual.*
Fusion Thesis: *Any fusion of individuals is an individual.*
Corollary: *Singletons are atoms.*

Besides a deduction of the Second Thesis from Division and Priority and Fusion, Lewis offers various motivating considerations and heuristic arguments in favor of these theses, but the justification of some ultimately remains largely pragmatic: The assumptions together lead to a powerful and attractive theory.

It is a consequence of these theses and mereology that the principle of the universality of set theory, according to which any condition determines a class whose members are all only those things for which the condition holds, barring proper classes, must fail, at least assuming for nontriviality that there is at least one individual and therefore at least one class. For that principle implies a strengthened version of division according to which there are only individuals and classes, whereas the fusion of an individual with a class cannot be a class by the Second Thesis, since it has an individual as a part, and cannot be an individual by the Priority Thesis, since it has a class as a part. Thus mereology and set theory cannot both be universal, and Lewis opts for mereology.

In practice, Lewis generally ignores fusions of individuals and classes, in effect tacitly assuming one is quantifying only over individuals and classes. For the most part, he also in effect assumes one is quantifying only over atoms (including singletons) and fusions thereof (including classes), to the exclusion of what he calls atomless gunk, if such there be. He sketches an adaptation of his results in the case where there is gunk. (If *all* there is is gunk, the assumption that there are lots of non-overlapping globs of it will do in place of the assumption that there are lots of atoms in the overall construction.)

29.2 Objections to Singletons

Mereology was founded in Poland by the nominalist Stanisław Leśniewski and popularized in the West by the nominalist Nelson Goodman precisely as a partial *substitute* for the set theory that they as nominalists rejected. Though Lewis is, by contrast, anything but a nominalist, still he does very much sympathize with traditional nominalistically inclined mereologists' complaints about set theory.

The background is as follows. Cantor's work on trigonometric series led him to move from thinking of the points (plural) where a function misbehaves to thinking of the set (singular) of points of misbehavior as a single object, to which operations can be applied, notably the operation of throwing away isolated points. Repeated application of this operation may have the result that only one point remains or none at all, and so it is a simple and natural step to admit singleton or unit sets and a null or empty set as "ideal elements" or limiting or degenerate cases, though in fact singletons and the empty set only really came to play an important role in set theory with Zermelo's axiomatization of the subject.

There is a certain kind of philosopher addicted to quibbling and querulous objections to mathematicians' habitual practice of rounding out systems of entities by positing "ideal objects," and more generally of counting in limiting or degenerate cases. Lewis gives a good impersonation of such a philosopher in some of his remarks on set theory, when he observes that Cantor's definition of set, which in free translation runs "any collection into a whole of definite, well-distinguished sensible or intelligible objects," is difficult to reconcile with reckoning in singletons, and impossible to reconcile with reckoning in an empty set.

Lewis considers a counterargument along the following lines. Suppose we have a stamp collection, complete with a catalog, and then, finding ourselves in reduced circumstances, have to begin selling off our stamps, deleting their listings from the catalog. This may go on until we have only one stamp left or none at all, leaving us with a singleton collection or an empty one. We need not even have a physical collection, but just a catalog, if we switch the example from stamp collections to Facebook selections of "favorites" in one or another category, if as fans we are fickle. Lewis rejects the argument on the grounds that all this talk of "collecting" is merely metaphorical. Well, of course it is: Cantor's notion of *set* was new – he did not suppose that what appears on the surface to be plural talk of points was deep down really singular talk of point-sets all along, and that he was merely making explicit what was already implicit – and a new notion can only be introduced by heuristic metaphors.

One would hardly expect such quibbling over minor ontological assumptions from, of all persons, David Lewis, he of the incredulous-stare-inducing ontology of real, concrete possible worlds. And the querulous objection is in any case pointless, since set theory with the limiting or degenerate cases can easily be interpreted in set theory without them. (As Lewis surely knew, at least by the time of MM. See Hazen 1991.) Indeed, assuming the existence of at least two individuals a and b, the "pure" sets of ordinary set theory can be mapped one-to-one onto those sets x such that x itself, all sets that

are elements of x, all sets that are elements of elements of x, and so on, have among their elements a, b, and *no other* individuals.

The real objection lies elsewhere. Nominalists have traditionally objected less to the part of Cantor's definition quoted so far, than to the additional clause "which are called the *elements* of the set," with its implication that set-formation is less a process of merger, like that by which Italy was formed from various minor states, than a process of federation, by which 13 colonies became the United States. The implication, to be more explicit, is that even after the many have been collected together into a one, it is still discernible *which* many they were: that just as the set is determined by its elements, so also the elements are determined by the set. Mereological fusion, by contrast, obliterates the separate identities of the fused: A single whole can be taken apart in many ways, and there is no one, canonical way of taking it apart of which it can be said that the genuine parts of which it is composed are just those pieces into which it is disassembled when taken apart in that way and no other.

Given Lewis's theses, to be an element of a set or member of class is just to have a singleton that is a part thereof. Grant the notion of singleton, and you have granted the notion of element or member, and the traditionally objectionable part of Cantor's definition. *That* is the real source of Lewis's objection to singletons, or rather, that together with the observation that many of the categories of metaphysics in the Australian style do not apply in any obvious way to singletons. (Is the relation of a singleton to its single member and internal or an external relation?) But though Lewis grouses and kvetches about such matters almost as much as a Leśniewskian or a Goodmanian might, to the point that one is expecting his discussion to issue in a proposal that if not literally nominalist would at least be in spirit nominalistic, on the contrary he ends by affirming that we must accept set theory *like it or not*.

For in the most memorable passage in all this material (PC 59; MM 15), Lewis writes that he laughs to think how *presumptuous* it would be to reject mathematics for philosophical reasons, and goes on to review the "great discoveries" of philosophy in the past, beginning with the proof of the impossibility of motion. This is perhaps a bit unfair, in that natural science after all emerged from natural philosophy; but the point stands that it is comically immodest for the part of philosophy that is still struggling, and therefore still called "philosophy," to seek to "correct" the part of philosophy that has succeeded, and is now called "science," and especially for anything as soft as philosophy to seek to "correct" mathematics, the hardest of the hard sciences. At any rate, with a forceful profession of faith in mathematics – he heads the relevant section "Credo" and might almost have followed the Tertullianists in adding "quia absurdum" – Lewis renounces renunciation of mathematics.

He remains tempted not by renunciation but by reinterpretation of certain kind. The reinterpretation in question is generally known in the contemporary literature as "structuralism," though it goes back (strictly speaking only in the case of arithmetic, though that case is easily adaptable to others) to Benacerraf (1965), rather than to any Parisian theoretician fashionable during the 1960s, apart perhaps from a very tenuous link to Bourbaki. The "structuralist" idea, which for Lewis is inspired by his reading of Ramsey, would be this, that instead of accepting a specific singleton-forming function, of philosophically inscrutable nature, simply to posit that there exists at least one function having the properties orthodox set theory ascribes to the singleton function. Lewis fears that even this degree of departure from strict and literal acceptance of set theory might constitute an unacceptable philosophical revisionism, but he perhaps need not have worried so much, for there are historical precedents.

In the seventeenth and eighteenth centuries, leading mathematicians (among them Descartes and Newton) had a more or less definite idea what (positive) real numbers were: ratios of magnitudes, such as lengths. In the nineteenth century, however, mathematicians came to feel that this geometric conception of the continuum needed to be replaced by something more purely arithmetic, and the constructions of Dedekind (his "cuts") and Cantor (equivalence classes of Cauchy sequences)

eventually emerged. By the early twentieth century such constructions were beginning to appear in undergraduate textbooks. G.H. Hardy, in his Cambridge freshman calculus textbook *Pure Mathematics* (in the second edition of 1914 and all subsequent ones), expounds Dedekind's construction, and then remarks that alternatives are possible, and that no great importance should be attached to the particular form of definition he has just finished presenting. He formulates – and quotes Bertrand Russell as endorsing – the general principle that in mathematics it matters that our symbols should be susceptible to *some* interpretation, but that if several are possible, it does not matter which we choose. Hardy's principle would seem to be just as applicable to set theory as to the calculus, and if so one has it on very high mathematical authority that there is nothing objectionable in the course that so tempts Lewis. Let us, in any case, see what that course involves.

29.3 Protestations of Innocence

Lewis's mereological assumptions amount to what mathematicians call the theory of a *complete Boolean algebra*, or, ignoring gunk, a *complete* atomic *Boolean algebra*, consisting of a number of atoms – just how many will turn out to be the great issue – and arbitrary fusions thereof. For what mathematicians call completeness amounts to what mereologists call unrestricted composition, permitting arbitrary fusions. There is just one departure from the usual mathematical approach, the dropping of the assumption of a null item in the algebra, which is one instance where Lewis *does* object to introducing an "ideal element" to round out a system. The minor complications this course involves him in will be ignored in the exposition below.

By completeness or unrestricted composition, absolutely arbitrary unions or joins are possible, as are absolutely arbitrary intersections or meets except those that would turn out to be null, and a complement exists for anything except the universal item or fusion of all things, whose complement would be null. Lewis, by the way, calls that universal fusion "Reality," and idiosyncratically calls two things whose intersection would be null – in other words, two things that are non-overlapping or disjoint – "distinct," a word more normally used to mean non-identical.

Unrestricted composition yields fusions of scattered, heterogeneous parts – fusions critics have considered monstrous and mythical. To his general defense of such chimeras (Lewis 1986), Lewis in PC adds a claim of a kind of ontological innocence, comparable to that of first-order logic, for mereology. The claim of ontological innocence is largely based on a variant of Donald Baxter's thesis of "composition as identity," the claim that when one thing is the fusion of many things, "They are it and it is them." (Though, as Hazen points out, one could believe that *when a fusion of many things exists* it is identical with the many things, without believing that such fusions exist unrestrictedly.)

Lewis takes this identity thesis somewhat less than literally, claiming that the relation of things to the fusion thereof is, though not strictly speaking identity, at any rate "analogous" to identity. It is difficult, however, to see how anything less than literal identity could suffice for ontological innocence. It is difficult, also, to see how Lewis can be acquitted of question-begging when he argues that one respect in which there is analogy is in ontological innocence.

For Lewis, the plural includes the singular: he does not object to the kind of counting in of limiting or degenerate cases involved in reading "there are some things" as "there are one or more things" (though he would object to reading it as "there are zero or more things"). The relation he finds analogous to identity amounts to the relation some things, the *x*s, bear to other things, the *y*s, just in case (i) the fusion of the *x*s is identical with the fusion of the *y*s, and (ii) either there is just one single *x* or there is just one single *y* or both. This includes identity of the usual kind, between one single *x* and one single *y*, as a special case.

Apart from symmetry, however, the relation in question lacks the usual formal properties of identity. It is not reflexive, since some two or more things, the *x*s, never stand in this relation to themselves, or to any other two or more things, the *y*s. And though it is transitive in the sense that when a single *x* bears this relation to some *y*s and those *y*s bear the same relation to a single thing *z*, then *x* is identical with *z*, it is intransitive in the sense that even when some two or more things, the *x*s, bear this relation to as single thing *y* and this *y* bears the same relation to some two or more things, the *z*s, the *x*s still do not bear this relation to the *z*s.

Nor need the *x*s then be identical with the *z*s as plural identity is usual understood. For the usual understanding requires that each single thing among the *x*s be also among the *z*s, whereas the eight ranks of a chessboard bear the Lewis relation to that chessboard, and the chessboard bears the Lewis relation to its eight files, while the ranks are not the files. Above all, as Lewis acknowledges, the indiscernibility of identicals fails utterly for plural things and their single fusions, since they are many while it is one. In our example, the ranks are horizontal while the files are vertical; and though there are eight ranks and eight files, there are 64 squares, whose fusion is again the same old chessboard.

It may be that in some sense the fusion is nothing over and above the things it is the fusion of, as Lewis asserts; but the things seem to be something over and above their fusion, consisting of that fusion plus a particular mode of division. Needless to say, the "plus" here is not *mereological* summation or fusion. In subsequent discussion on "composition as identity" – see Sutton (2008) for an overview – critics have outnumbered defenders. Yet even on this, his weakest point, Lewis's discussion, in his inimitable style, remains well worth reading.

29.4 Background on Mereoplethynticology

In the mathematical literature, the completeness assumption for a Boolean algebra is generally stated in second-order terms: For any class of things, there is a fusion thereof. In the context of Lewis's project this will not do. Instead he draws on the resources of Boolos's plural quantification, which the present writer has elsewhere called *plethynticology*, and says simply that for any things (plural) there is a fusion (singular) thereof.

Technically, the only substantive assumption of plethynticology is that for any condition (that holds of at least one thing) there are some things such that they are precisely the things for which the conditions hold. (And which things are these? Those for which the condition holds, obviously.) Philosophically, Lewis joins Boolos in emphatically rejecting the suggestion that plural quantification over things secretly involves singular quantification over pluralities of things, construed as single things of some extraordinary set- or class-like kind, and inveighs at some length over such "singularism." For a fuller polemic even than Lewis's against that misguided view, see McKay (2006).

The effect is that Lewis has the resources to interpret *monadic third-order logic*. In the terminology of Hazen (1997) for reading the formalism of that logic, a first-order variable *x* is said to range over *individuals*, a second-order variable X is said to range over *species*, and a third-order variable *X* is said to range over *genera*. For Lewis, the "individuals" are atoms, the "species" are fusions of atoms, and singular quantification over "genera" would be replaced by plural quantification over such fusions of atoms. Likewise *couples*, or species with just two individuals belonging to them, for Lewis are *diatoms*, or fusions of just two atoms. Having noted that this is Lewis's official interpretation, for the purely formal developments to follow it will be convenient to fall in with Hazen's terminology, leaving the reader to work out the translation.

Following Lewis (though speaking like Hazen), one would like at this point to formulate the assumption that there exists a function for which certain conditions hold, characteristic of the

singleton function, in terms of which we can define a binary relation on individuals for which certain other conditions hold, characteristic of the membership relation. But once we have gone structuralist, we may as well just assume the existence of such a binary relation. What we want is a binary relation such that the following hold:

(1) For any y only a few x stand in the given relation to y.
(2) For any small species X there is a unique individual y such that the individuals x standing in the given relation to y are precisely those belonging to X.

We can then call those y for which there is at least one x standing in the given relation to y our nonempty sets, and then call the relation in question set-membership or elementhood – that so doing may have the consequence that Julius Caesar turns out to be a set would have troubled Frege, but does not trouble Lewis, who does not count intuitions to the effect that sets are nonphysical and nonmental and nonspatial and nontemporal as official parts of the set theory he is trying to preserve – and try to see what further assumptions are needed to recover the usual axioms of first-order ZFC, and thence of second-order ZFC, letting fusions of many sets play the role of proper classes.

Note that we are going to need quantification over relations anyhow, in order to define what it is for a species to be small (there is no one-to-one correlation between its individuals and *all* individuals), or for the individuals for which some condition holds to be few (the species of such individuals is small). But while in monadic third-order logic we have variables x and X and X for individuals and species and genera, we have no variables for Ξ for binary relations on individuals. This brings us to the technical point on which there was to be progress between the first draft of PC and the published version including PC*: the simulation of quantification over binary relations Ξ on individuals. Two methods (a) and (b) of simulation were suggested in PC*, a hybrid (ab) between them was suggested in MM, while a new modification (b′) of (b) will be suggested below.

Note that in the framework as described so far, we do have available unordered pairs and therewith *symmetric relations*, genera with only couples belonging to them. Moreover there is a general result in mathematical logic (see Boolos, Burgess and Jeffrey 2007, §21.3) for coding any finite number of relations of any finite number of places by a single binary relation. Unfortunately, the method requires that one have available, in addition to the domain of individuals in which one is primarily interested, a further domain of auxiliary individuals, which is not an assumption likely to tempt Lewis.

All the alternatives, making do with a single domain of individuals, have in common that they involve the axiom of choice (henceforth AC). The use of AC in one way or another is inevitable, according to the results of Hazen (1997) – though in the same paper Hazen also indicated that special features of set theory would have made it possible for Lewis to formulate a structuralist version of *it* without general quantification over relations, and without AC.

29.5 Axioms of Choice

Method (a), developed in PC* from a suggestion of the present writer, involves also the assumption that there are infinitely many individuals. Of course, that will follow from orthodox set theory once we are in a position to formulate it; but it may be thought undesirable to presuppose it in the very formulation of our set-theory-generating assumptions. How AC comes in with method (a) is perhaps the only point worth elaborating before moving on to alternatives. The infamous Banach-Tarski paradox (according to which a ball can be cut up into finitely many pieces, and the pieces after

translation and rotation fit back together to form two balls each the size of the original) depends ultimately on a simple consequence of AC: that if a set is infinite, it can be decomposed into three non-overlapping pieces, any one of which is equinumerous with the union of the other two. That is the assumption that is used in method (a).

Method (b) begins with what is (as was subsequently recognized) an old trick of earlier set theorists (from the period before the Weiner-Kuratowski set-theoretic definition of ordered pair) that makes it possible within our framework to simulate quantification over linear orderings of individuals. Such an ordering can be represented by the genus to which belong all and only those species that are initial segments in the ordering. Such a linear ordering makes it possible, in a systematic way, to distinguish given any two individuals x and y, one that is *preferred* and one that is *spurned*. These are simply the one that comes earlier and the one that comes later in the linear order.

Hazen in effect combined these considerations with another observation, that given such a way of distinguishing a preferred from a spurned individual among any pair of individuals, an arbitrary binary relation can be represented by a species, that of those individuals that are self-related, together with two symmetric relations. One is the symmetric relation that holds between two individuals if and only if the preferred one stands to the spurned one in the given relation; the other is the symmetric relation that holds between two individuals if and only if the spurned one stands in the given relation to the preferred one. In this way, a quantifier ranging over relations can be simulated by three quantifiers, one ranging over species and the other two over genera representing symmetric relations. All this, however, is provided we assume there exists at least one linear ordering of all individuals. The existence of such a linear ordering is another consequence of the axiom of choice.

Now in a set-theoretic context there are numerous equivalents, and innumerable consequences, of AC. All have historically been to a degree controversial, but the most basic formulation of AC has had wide appeal. Gödel (1947) emphatically declared it to be "evident" (given that we clear our heads of any notion that the existence of sets somehow depends on our being able to define them). A number of early set theorists implicitly assumed it without even noticing that they were doing so. It asserts that for any set of non-empty, non-overlapping sets, there is a set that overlaps each of them in one and only one element (or in our context, for any genus of non-empty, non-overlapping species, there is a species that overlaps each of them in one and only one individual).

From this there follow the various equivalents and consequences, some of which (such as the Banach-Tarski result) have been declared "paradoxical" and "counterintuitive," and are admitted even by Gödel to be "surprising and unexpected." The assumption about the decomposability of any infinite set used in method (a), and about the linear orderability of any set used in method (b), are among such "surprising and unexpected" consequences, only to be accepted because they are derivable from the basic, "evident" version.

Unfortunately, though the derivations are, in the context of set theory, where we have ordered pairs freely available, easy enough to be included in undergraduate courses, nothing like those derivations is possible within the Lewis framework or monadic third-order logic. There is no difficulty in *stating* the basic, "evident" form of the axiom, but the derivation of the *useful* consequences, including those used in methods (a) and (b), requires that we have ordered pairs or surrogates for them available already, which we do not.

This suggests rethinking the issue, and the first thing to notice is a point already incorporated into the exposition above, that Hazen's method (b) really only requires that it should be possible, given any two individuals, to label one preferred and the other spurned, in some systematic way. The preferences need not cohere into a linear preference order: Condorcet cycles, in which the preference goes to x over y and y over z but z over x, would be no obstacle. And there is a treatment of the axiom of choice that would make such a systematic choice of a preferred one out of any two immediately available.

The background is as follows. Initial opposition to AC stressed its *nonconstructive* character: Even in its basic, "evident" form it assumes the existence of a set satisfying certain conditions without giving an explicit or implicit definition of any such set. It was soon realized, however, that the nonconstructive character of orthodox mathematics by no means begins with set theory, but is to be found already in its basic logic. Classical first-order logic already assumes the validity of the scheme $\exists x(\exists y A(y) \rightarrow A(x))$ for any condition expressible by a formula A, without giving any notion of how to find such an x. (The proof is simply that if there is at least one y such that $A(y)$, then any such y will do for x, while if there is no such y, then anything at all will do for x. As the heresiarch Brouwer, founder of intuitionism, noted, the nonconstructivity comes in the principle of *excluded middle*, according to which either there is or there isn't such a y.) Something like this observation lies behind David Hilbert's proposal to introduce his ε-symbol. Given a formula A, with this symbol we can form a term $\varepsilon x A(x)$ substitutable for individual variables, and the assumption is that $\exists y A(y) \rightarrow A(\varepsilon x A(x))$. (The suggestion of Hilbert that one could in fact then *define* quantification $\exists y A(y)$ to *mean* $A(\varepsilon x A(x))$ will here be ignored.)

Mathematicians have a way of reasoning in which, having proved or assumed the existence of at least one x such that $A(x)$ they allow themselves to introduce a term, saying, "Let a be such an x," and more generally, having proved or assumed the existence for each y of at least one x such that $A(y, x)$, to introduce a notation, saying, "For each y, let a_y be such an x." The Hilbert ε-symbol is a formal representation of this way of reasoning: a is $\varepsilon x A(x)$, and a_y is $\varepsilon x A(y, x)$. The usual formal axiomatics of set theory does not allow such notation to be used in defining sets (that is, in instances of the separation and replacement scheme); if it is allowed, AC becomes deducible, and does not need to be assumed as an axiom. The Hilbert ε-symbol in effect builds AC into the background logic.

The ε-symbol is most useful if one makes the following additional assumption of *extensionality*:

$$\forall x(A(x) \leftrightarrow B(x)) \rightarrow \varepsilon x A(x) = \varepsilon x B(x)$$

In a context where we have species in addition to individuals, starting from an ε-symbol for which we do *not* have extensionality in this sense, we can get one for which we *do*, by defining $\varepsilon^* x A(x)$ to be the y such that

$$\exists z(\text{z is the species of all } x \text{ such that } A(x) \,\&\, y = \varepsilon x(x \text{ is an individual belonging to z}))$$

So we might as well assume extensionality from the beginning. In that case we easily get, for any a and b, a preferred one of the two, namely, $\varepsilon x(x = a \vee x = b)$. Incorporating the Hilbert ε-symbol, embodying the "evident" form of AC, into the background logic, and thereafter following Hazen, constitutes method (b′) for simulating quantification over relations. It provides perhaps the neatest way to surmount the obstacle to Lewis's program.

29.6 Multitudes of Individuals

The reason Frege's *Grundgesetze* system is inconsistent is that his assumption that every "concept" has an "extension" (in our context, the assumption that there is a one-to-one function from *all* species to individuals) leads to the Russell paradox. Boolos (1989) considers the fallback assumption, begotten by Cantor and baptized "limitation of size" by Russell, that only "small" concepts have "extensions." Restated in Hazen's terminology rather than Frege's, the *fallback assumption* is that there are no more small species than individuals.

This is another informal wording of the same assumption displayed in formulations (i) and (ii) of section 29.4. It represents one way of assuming that there are *many* individuals. In particular, it implies there is an infinity of them, assuming we have at least three to get going with. For even disallowing a null or empty species, given n individuals there are $2^n - 2$ small species, and therefore by the fallback assumption at least that many individuals: 3 gets us $2^3 - 2 = 6$, 6 gets us $2^6 - 2 = 62$, 62 gets us $2^{62} - 2 > 4{,}000{,}000{,}000{,}000{,}000{,}000$, and so on.

The very formulation of the fallback assumption requires quantification over relations on individuals, but we have just seen that there are several ways to provide a simulacrum of that, through appropriate applications of AC. Given this fallback assumption, the "structuralist" idea is then that while we allow ourselves to write, as Boolos does, as if we had in mind some fixed, specific membership relation $\in$, our real understanding is that all our assertions are tacit generalizations about *any* relation of the right kind.

With this understanding, Boolos (1989) in effect tells us that we automatically get, exploiting known results pertaining to von Neumann's old approach to axiomatizing set theory, all of second-order ZFC or MK (indeed, as Hazens reminds us, in the Lewis framework through plural quantification over class-surrogates, third-order ZFC, or second-order MK) *except* the axioms of infinity and power set (at any rate assuming there are at least three individuals, so that two is few, an assumption needed to get the usual pairing axiom). Some of the derivations are a bit tricky: the axiom of unions requires a slight modification of a trick due to Levy, the axioms of choice and foundation a couple of tricks due to von Neumann.

Lewis, of course, with his "Credo," wants all of orthodox set theory, including the infinity and power axioms, but what one needs to get these are extra assumptions easily formulable within his framework and ours, and formulable as just assumptions to the effect that there are lots of individuals. For infinity, we need merely that there should be some *small* species of individuals that is infinite; and, for power, the assumption that for every small species, its subspecies are fewer than there are individuals. These are both, like the fallback assumption itself, in effect assertions about *how many* individuals there are, or in Lewis's terms, about "the size of Reality."

Set theorists have considered ever more daring extensions of the usual axioms, but these two generally take the form of "higher axioms of infinity" or "large cardinal hypotheses," which mostly lend themselves to reformulation within the frameworks under consideration as well. The mildest of these (the existence of arbitrarily large inaccessible cardinals) has even, according to McLarty (2010), been used (though really only as a dispensable convenience) in certain category-theoretic work whose results have been appealed to in works themselves in turn appealed to in the first published proof of Fermat's theorem.

That much and a bit more in the way of large cardinal theory is neatly incorporated in an extension of second-order ZFC or MK known as Bernays set theory. (See Burgess 2004 for a formulation based on plural quantification.) Bernays manages to replace most of the usual axioms of set theory by a single assumption, a so-called *reflection principle*, to the effect that there are "indescribably many" individuals. What this means is that for any statement Θ that holds of individuals and species there is a small species such that Θ continues to hold when individual variables are restricted to individuals belonging to that species, and species variables to subspecies of that species.

According to reflection, intuitively speaking, any attempt to describe how many individuals there are turns out, if true, to be an understatement, which would be equally true if one were speaking not of *all* individuals, but only of the *few* individuals belonging to some small species. Thus the statement that there exists at least one individual, in other words, that the domain over which the individual variables range is nonempty, which by courtesy is considered a principle of logic, upon reflection yields the consequence that there is a *small* species to which at least one individual belongs, and hence that there are at least *two* individuals.

Reflecting on *that* assertion yields *three*, and reflecting on *that* assertion yields *four*, and so on. Thus we have infinitely many individuals – even *without* the fallback assumption or the assumption we have at least three individuals to get going with – and reflecting on *that* conclusion yields a small species that is infinite, just what is needed for the set-theoretic infinity axiom. We do still need to make, in addition to the assumption of reflection, the fallback assumption that there are no more small species than individuals, but application of an appropriate version of reflection to *that* assumption turns out to yield just what is needed for the set-theoretic power axiom.

In the end, a framework of (i) first-order logic enriched by the Hilbert ε-symbol and (ii) Boolos's plural quantification and (iii) the assumption that the parthood relation of mereology obeys the laws of a complete atomic Boolean algebra, leaves us needing only (to revert from Hazen's to Lewis's terminology) (iv) the fallback or limitation-of-size assumption that there are no more small fusions of atoms than there are atoms, and (v) the reflection principle that there are indescribably many atoms, to get all of standard set theory, and therewith all of classical mathematics, plus the more modest of the large cardinal hypotheses to boot. The special ontological assumptions (iv) and (v) of mathematics, beyond the framework (i)–(iii) indicated, amount in the simplest terms to just to this, that there are a lot of atoms of whatever nature around (to serve as sets, or, if you please, codes or surrogates for sets). This is not the place to go into technical details, and so I draw my sketch to a close. Suffice it to say that we have here, in the mereological treatment of set theory, when all the details of the outline sketched above are filled in, a very elegant application of Ludovician metaphysics.

References

Benacerraf, P. (1965). What numbers could not be. *Philosophical Review*, 74, 47–73.

Boolos, G. (1971). The iterative conception of set. *Journal of Philosophy*, 68, 215–31.

Boolos, G. (1984). To be is to be the value of a variable (or some values of some variables). *Journal of Philosophy*, 81, 430–49.

Boolos, G. (1985). Nominalist platonism. *Philosophical Review*, 94, 327–44.

Boolos, G. (1989). Iteration again. *Philosophical Topics*, 17, 5–21.

Boolos, G., Burgess, J.P. & Jeffrey, R. (2007). *Computability and logic*, 5th ed. Cambridge: Cambridge University Press.

Burgess, J.P. (2004). *E pluribus unum*: plural logic and set theory. *Philosophia Mathematica*, 12, 193–221.

Burgess, J.P., Hazen, A.P. & Lewis, D. (1991). Appendix on pairing. In Lewis (1991), pp. 121–49.

Gödel, K. (1947). What is Cantor's continuum problem? *Mathematical Monthly*, 9, 515–25.

Hazen, A.P. (1991). Small sets. *Philosophical Studies*, 63, 119–23.

Hazen, A.P. (1997). Relations in monadic third-order logic. *Journal of Philosophical Logic*, 26, 619–28.

Lewis, D. (1986). *On the plurality of worlds*. Oxford: Blackwell.

Lewis, D. (1991). *Parts of classes*. Oxford: Blackwell.

Lewis, D. (1993). Mathematics is megethology. *Philosophia Mathematica*, 1, 3–23.

McKay, T. (2006). *Plural predication*. Oxford: Oxford University Press.

McLarty, C. (2010). What does it take to prove Fermat's last theorem? Grothendieck and the logic of number theory. *Bulletin of Symbolic Logic*, 16, 359–77.

Sutton, C. (2008). Composition as identity: a resource page, http://www.people.vcu.edu/∼csutton2/composition.html, accessed June 6, 2014.

Varzi, A. (2011). Mereology. In E.N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2011 edition), http://plato.stanford.edu/archives/spr2011/entries/mereology/, accessed June 6, 2014.

Part V

Epistemology and Mind

30

Lewis on Knowledge Ascriptions

JONATHAN SCHAFFER

David Lewis's primary contribution to the theory of knowledge is his account of knowledge ascription, which integrates an elegant version of relevant alternatives theory with a detailed version of contextualism. His account is prefigured in his discussion of accommodation in "Scorekeeping in a Language Game" (1979), and forms the central topic of his "Elusive Knowledge" (1996). I will review his account (section 30.1), and then discuss some epistemic (section 30.2) and semantic (section 30.3) questions arising. This will be a friendly critique: I am sympathetic with the outlines of Lewis's contextualist relevant alternatives theory, but would amend some details.

Before proceeding, I should mention that Lewis offers discussions bearing on knowledge in various places, including:

- the treatment of conventions as common knowledge, provided in *Convention* (1969);
- the semantics of 'whether'-clauses, including their behavior in 'know whether' constructions, found in "Whether Report" (1982);
- the concern that modal realism entails modal skepticism, discussed in *On the Plurality of Worlds* (1986);
- the claim that knowing what it is like to have an experience does not consist in having information but only in having abilities, defended in "What Experience Teaches" (1988a); and
- the argument that we cannot know which properties occupy the theoretical roles of our best theory, presented in "Ramseyan Humility" (2009).

In these discussions Lewis tends to use the notion of knowledge in an intuitive way, and it is the burden of his contextualist relevant alternatives theory to fit this intuitive usage. I leave open the exegetical question of whether Lewis's theory always fits his usage.[1]

I should also mention that Lewis offers various extended discussions of "formal epistemology," including:

- the treatment of objective chance via the Principal Principle, articulated in "Subjectivist's Guide to Objective Chance" (1980);

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

- the argument that treating desire as belief conflicts with standard decision theory, developed in "Desire as Belief" (1988b);
- the diachronic Dutch book argument for updating by Bayesian conditionalization, motivated in "Why Conditionalize?" (1999); and
- the argument that Sleeping Beauty ought to assign a credence of .5 to heads when she wakes, given in "Sleeping Beauty: Reply to Elga" (2001).

I leave open what if any connections there might be between Lewis's treatment of formal epistemology and his treatment of "traditional" matters of knowledge and its ascription, save to note that he (1996, 553) characterizes our concept of knowledge as a "handy but humble approximation" which only provides "a very sloppy way of conveying very incomplete information about the elimination of possibilities."

One might think that Lewis, by speaking of our concept of knowledge as a "handy but humble approximation," means to marginalize the topic. But he (1996, 553) immediately adds that the concept may yet be "indispensable in practice," and he evidently deems it worthy of detailed discussion. Indeed a great deal of his overall philosophical energy went into developing philosophically rigorous accounts of shifty but handy concepts such as counterfactuals and causation.[2] So I think that he regards knowledge as one of the many shifty but handy concepts for which a rigorous philosophical treatment is a difficult but necessary task.

30.1 Lewisian Knowledge Ascription

Lewis offers an account of knowledge ascription, which integrates relevant alternatives theory with contextualism. I begin with a review. This will involve presenting both his elegant version of relevant alternatives theory (section 30.1.1), and his detailed version of contextualism (section 30.1.2), and then displaying the combined account and its claimed virtues (section 30.1.3).

30.1.1 Lewisian Relevant Alternatives Theory

Lewis's account begs to be misunderstood. From his title ("Elusive Knowledge") up to the very end of his paper – including the "official statement" of his view – Lewis speaks *as if* he were presenting a theory of knowledge. Only at the end does he (1996, 566–7) mention that he has offered a simplified and potentially misleading presentation. But he still offers no corrected statement, instead trusting the reader who senses something amiss to appreciate how his account could have been stated in a more cumbersome but correct form, via semantic ascent. This is a lot of trust to place in the reader. So – if only to help the reader repay this trust – it will prove useful to state Lewis's account correctly.

Thus the first point to clarify is that Lewis (despite simplified statements as if otherwise) is not really presenting an account of *knowledge*, but rather of *knowledge ascription*. He is not really saying when *s* knows that *p*, but rather when a sentence of the form '*s* knows that *p*' is true relative to a context *c*. These are connected but distinct topics.[3] As such he is *not* really saying that knowledge is elusive, or is destroyed by doing epistemology, or anything like that (despite seeming to say just these things). Properly understood, he is really saying something metalinguistic, about the contexts in which knowledge ascriptions come out true.

With this simplification firmly in mind, one can approach Lewis's (1996, 551) official statement: "Subject *S knows* proposition *P* iff *P* holds in every possibility left uneliminated by *S*'s evidence; equivalently, if *S*'s evidence eliminates every possibility in which not-*P*." The context dependence is

said to arise from the context dependence of the domain associated with 'every', which he (1996, 554) incorporates as: "*S* knows that *P* iff *S*'s evidence eliminates every possibility in which not-*P* – Psst! – except for those possibilities that we are properly ignoring." Semantically ascending, one reaches the more cumbersome metalinguistic claim that correctly states Lewis's account:

> **Lewisian relevant alternatives theory:** A sentence of the form '*s* knows that *p*' is true in context *c* iff *s*'s evidence eliminates every not-*p* possibility relevant in *c*.

Lewisian relevant alternatives theory is distinctive in at least two main respects, the first of which – though obscured in Lewis's simplified presentation – is its metalinguistic character.[4] This proves crucial for integration with contextualism. With an object language account, contextualism would lead straight to contradiction, since one could derive both *s* knows that *p* (from a "favorable" context) and then its negation (from an "unfavorable" context). A metalinguistic account only allows one to derive that a given sentence (of the form '*s* knows that *p*') is true in one context and false in another, and that is not only consistent and unproblematic, but just what one expects from contextually sensitive discourse. (By way of comparison, it is unproblematic and indeed expected that – due to the context sensitivity of 'today' – the sentence 'today is Friday' is true in a Friday context but false in a Monday context.)

The metalinguistic character of *Lewisian relevant alternatives theory* allows for contextualism but does not demand it. Context invariance is equally allowed insofar as nothing yet has been said about the determinants of relevance (Lewis's account of relevance will be discussed in section 30.1.2), and so nothing yet rules out the prospect that exactly the same relevant alternatives might be generated in every single context, given a fixed subject *s* and proposition *p*. Thus it is strictly compatible with *Lewisian relevant alternatives theory* to hold that, for any subject *s* and proposition *p*, the relevant alternatives in every context are (i) every not-*p* possibility which is unrestrictedly logically possible (Unger 1975), or (ii) every not-*p* possibility with nonzero objective chance in *s*'s situation (Dretske 1991, 192), or even (iii) every not-*p* possibility which is practically relevant to *s* (Hawthorne 2004, 158–61; Stanley 2005, 85). Thus *Lewisian relevant alternatives theory* is usefully neutral on contextualism.

A second main respect in which *Lewisian relevant alternatives theory* is distinctive is in its elegance, in making the elimination of the relevant alternatives the *one and only* requirement. No other requirements – such as justification, truth, belief, or basing – are imposed.[5] Instead Lewis uses his account of relevance – to which I now turn – to simulate some of these requirements, in distinctive ways.

30.1.2 Lewisian Contextualism

Lewis combines the elegance of *Lewisian relevant alternatives theory* with a detailed contextualist account of the determinants of relevance. Indeed a large portion (pp. 554–61) of "Elusive Knowledge" consists in discussing rules for determining the relevance of a possibility at a context. Lewis offers seven such rules:

Actuality: the possibility that actually obtains for the subject is always relevant;
Belief: any possibility that the subject believes or ought to believe obtains is always relevant;
Resemblance: any possibility that saliently resembles a relevant possibility (made relevant by any rule other than *Resemblance* itself) is always relevant;
Reliability: possibilities concerning errors in reliable processes (such as perception, memory, and testimony) are defeasibly irrelevant;
Method: possibilities concerning errors in sampling and in abduction are defeasibly irrelevant;

Conservatism: possibilities that are conventionally ignored are defeasibly irrelevant; and
Attention: any possibility explicitly under discussion is always relevant.

Putting these seven rules together yields:

> **Lewisian contextualism:** The relevance of a possibility at a context is determined by all and only the rules of *Actuality*, *Belief*, *Resemblance*, *Reliability*, *Method*, *Conservatism*, and *Attention*.

Lewis's rules can be usefully divided in several ways, of which I will mention three. First, there are factors that push towards expanding the relevant possibilities (*Actuality*, *Belief*, *Resemblance*, *Attention*), and factors that push towards contracting them (*Reliability*, *Method*, *Conservatism*). Second, there are factors that push with irresistible force (*Actuality*, *Belief*, *Resemblance*, *Attention*), and those that push with resistible force possibly overcome by a push in the other direction (*Reliability*, *Method*, *Conservatism*).[6] Though – as will emerge in section 30.3 – Lewis largely undoes this second distinction at the end of his paper, allowing that all the factors he cites are resistible. Third – and most importantly for present purposes – there are invariant factors that are fixed across attributors (*Actuality*, *Belief*, *Reliability*, *Method*, *Conservatism*), and contextual factors that can vary across attributors (*Resemblance*, *Attention*). It is the postulation of these latter factors that make Lewis's account contextualist. Lewis's rules minus *Resemblance* and *Attention* do not allow for any two contexts to differ in their relevant possibilities, given a fixed subject *s* and a fixed proposition *p*.

Lewisian contextualism is distinctive in at least two main respects, the first of which is its attention to detail. Previous relevant alternatives theorists – whether contextualists or not – had left the notion of relevance at a largely intuitive level, leading previous commentators such as Sosa (1986, 585) to speak of relevance as "unacceptably occult." While *Lewisian contextualism* leaves some details open,[7] it achieves a level of precision that might be thought appropriate to the topic, and at any rate clearly surpasses anything found before.

A second main respect in which *Lewisian contextualism* is distinctive is its treatment of factors like justification, truth, and belief (this is the counterpart of the elegance of *Lewisian relevant alternative theory*, which did without mention of these matters). The treatment of truth turns out to be distinctive only in implementation, not in substance. Lewis imposes the usual requirement that the proposition *p* be true, just indirectly via the rules of relevance instead of directly via a condition on knowledge ascriptions.[8] But the treatments of justification and belief are distinctive in substantive ways. For Lewis (1996, 551) denies that justification is necessary for knowledge, and he (1996, 556) likewise denies that belief is necessary for knowledge.[9] Instead his *Belief* rule is supposed to function as an improved replacement for these requirements, giving belief and justification a minor role in generating relevance (as Lewis (1996, 556) notes: "This is the only place where belief and justification enter my story"), while still allowing for a true knowledge ascription in the absence of either belief or justification, given sufficient evidence for ruling out whatever alternatives got generated.

30.1.3 Lewisian Knowledge Ascription

Lewis's full theory of knowledge ascription is then the result of conjoining *Lewisian relevant alternatives theory* with *Lewisian contextualism*:

> **Lewisian knowledge ascription:** A sentence of the form '*s* knows that *p*' is true in context *c* iff *s*'s evidence eliminates every not-*P* possibility relevant in *c*, where the relevance of a possibility in *c* is determined by all and only the rules of *Actuality*, *Belief*, *Resemblance*, *Reliability*, *Method*, *Conservatism*, and *Attention*.

To illustrate, consider the following sentence:

(1) Ann knows that there is a goldfinch in the garden.

By *Lewisian knowledge ascription*, (1) is true in a given context *c* iff Ann's evidence eliminates every possibility relevant in *c* in which it is not the case that there is a goldfinch in the garden. For instance, suppose that in context c_{easy} the only relevant possibility in which it is not the case that there is a goldfinch in the garden is the possibility in which there is a raven in the garden instead, and suppose that Ann's evidence eliminates the raven possibility. Then (1) is true in c_{easy}. (One might think of c_{easy} as a context in which the question under discussion is whether there is a goldfinch or a raven in the garden, and one might think of Ann as someone in position to answer this easy question.) But suppose that in context c_{hard} there is a relevant possibility in which there is a canary in the garden instead, which Ann's evidence does not eliminate. Then (1) comes out false in c_{hard}. (One might think of c_{hard} as a context in which the question under discussion is whether there is a goldfinch or a canary in the garden, and one might think of Ann as someone not in position to answer this hard question.)

It is worth mentioning that *Lewisian relevant alternatives theory* and *Lewisian contextualism* are independent. (This is why I have presented *Lewisian knowledge ascription* in conjunctive form, and why I will split the questions arising into two groups, in sections 30.2–30.3.) One can accept *Lewisian relevant alternatives theory* but reject *Lewisian contextualism*, simply by endorsing different rules of relevance. As mentioned above, one could even endorse rules that do not include any factors that can vary across contributors, thus rendering the resulting package invariantist. And likewise one can accept *Lewisian contextualism* but reject *Lewisian relevant alternatives*, simply by having a different account of the truth conditions for knowledge ascriptions. One could for instance add in a belief and/or justification condition. One could even endorse a semantics for knowledge ascriptions in which the relevance of alternatives played no role whatsoever, and relocate Lewis's rules of relevance to the realm of conversational pragmatics.[10]

Lewis then claims several virtues for the package that is *Lewisian knowledge ascription*, including two already mentioned: improving the connection between knowledge, belief, and justification (by treating the latter two as partial determinants of relevance rather than requirements for the truth of a knowledge ascription), and fitting our shifty intuitions for knowledge ascriptions (as seen in the treatment of (1) at c_{easy} and c_{hard}). With respect to fitting our shifty intuitions, subsequent empirical work looks to vindicate Lewis's intuitive idea that our intuitions about knowledge ascriptions are in fact sensitive to the contextually relevant alternatives.[11] (Though of course the invariantist can try to account for these intuitions in terms of conversational pragmatics or performance errors, instead of via the semantics.[12])

But the main virtue which Lewis claims for *Lewisian knowledge ascription* is that of reconciling infallibilism with anti-skepticism. Thus Lewis (1996, 550) opens with a dilemma: "[W]e know a lot; knowledge must be infallible; yet we have fallible knowledge or none (or next to none). We are caught between the rock of fallibilism and the whirlpool of skepticism. Both are mad!" His account is primarily motivated as resolving this dilemma. The resolution is to grant the infallibilist her principle that knowledge requires the elimination of "every" relevant alternative, but understand this principle in a context-dependent way so that skeptical doubts (such as being a brain-in-a-vat) are not ordinarily relevant. Ordinarily such doubts do not come up for consideration at all, and so – in such ordinary contexts – knowledge ascriptions can still come out true. But in extraordinary contexts when such doubts come unto consideration, knowledge ascriptions go false, as he (1996, 551) explains:

> Maybe epistemology is the culprit. Maybe this extraordinary pastime robs us of our knowledge. Maybe we do know a lot in daily life; but maybe when we look hard at our knowledge, it goes away. But only when we look at it harder than the sane ever do in daily life; only when we let our paranoid fantasies rip. That is when we are forced to admit that there always are uneliminated possibilities of error, so that we have fallible knowledge or none.

(Though the reader should keep in mind that when Lewis speaks of epistemology as robbing us "of our knowledge," he really means something metalinguistic. He means that engagement in epistemology can create a type of context in which 'know' expresses an extremely demanding relation.)

Putting this together: Lewis offers an account of knowledge ascription – *Lewisian knowledge ascription* – which conjoins a distinctively metalinguistic and elegant form of relevant alternatives theory – *Lewisian relevant alternatives theory* – with a distinctively detailed and powerful form of contextualism – *Lewisian contextualism*. Not for nothing has "Elusive Knowledge" earned a central place in the contemporary contextualist canon alongside other pioneering work such as that of Cohen (1988) and DeRose (1995), and inspired a range of subsequent approaches including those of Schaffer (2005), Blome-Tillmann (2009), and Ichikawa (2011).

30.2 Epistemic Questions

So far I have spelled out Lewis's theory via *Lewisian knowledge ascription*, which conjoins *Lewisian relevant alternatives theory* and *Lewisian contextualism*. But the theory is in some respects incomplete, and in other respects problematic. Without pretense of covering every possible concern, let me just pose some questions that any friend of contextualist relevant alternatives theory needs to address. I will begin with three questions which might be thought of as *epistemic questions*, insofar as they primarily concern *Lewisian relevant alternatives theory*. (The *semantic questions*, to be discussed in section 30.3, will be those that primarily concern *Lewisian contextualism*.)

30.2.1 Hyperintensionality

One obvious question is whether a framework of possibilities (the sort of things that might or might not be relevant, on Lewis's approach) provides a sufficiently *fine-grained* framework for a hyperintensional topic like epistemology. Assuming that $5 + 7 = 12$ and Fermat's Last Theorem are both necessary truths holding at every possibility, *Lewisian Knowledge* entails that '*s* knows that $5 + 7 = 12$' and '*s* knows that Fermat's Last Theorem holds' are both true, for any context *c*. This is counterintuitive. It seems that in most contexts we would affirm, of a normal human adult Ann who is not a mathematician:

(2) Ann knows that $5 + 7 = 12$.

But we would reject:

(3) Ann knows that Fermat's Last Theorem holds.

So one might worry that *Lewisian relevant alternatives theory* is built on too coarse a framework.

Lewis (1996, 551–2) touches on the matter of hyperintensionality, accepting (in accord with his general views) that "there is only one necessary proposition" but allowing that this "known

proposition may go unrecognized when presented in impenetrable linguistic disguise, say as the proposition that every number is the sum of two primes." But he immediately adds (before moving on to other matters): "These problems of disguise shall not concern us here. Our topic is modal, not hyperintensional, epistemology."

Lewis's parting comment is odd, since epistemology is evidently a hyperintensional topic. There is no such topic as "modal epistemology." (Imagine someone whose account of knowledge failed to entail factivity, and who dismissed the worry on grounds that her topic was "nonfactive epistemology.") And Lewis's talk of disguise is odd as well, insofar as disguises are in no way integrated into his positive theory. (As it might have been if Lewis had taken knowledge to be a ternary relation between a subject, a proposition, and a guise.) So *Lewisian Knowledge* still entails that (2) is true iff (3) is true, for any context *c*. The seeming counterexample stands untouched.

Moreover, Lewis's talk of disguise is unhelpful in *third-person cases* in which the ascriber can "see through the disguise." For instance, if Beth is a math professor who has just carefully reviewed Wiles's proof of Fermat's Last Theorem, and Ann is a first-year student, it seems perfectly plausible that Beth might both affirm (2) and deny (3). Whatever is happening in this case, Beth (the ascriber) is clearly *not* failing to recognize the necessary proposition under an "impenetrable linguistic guise." After all, she has just carefully reviewed Wiles's proof.

Let me suggest an alternative answer on Lewis's behalf, which begins in a more concessionary mode: epistemology is a hyperintensional topic. But this alternative makes no further concessions, and instead notes that hyperintensionality is a problem for everyone (tu quoque). No one has yet developed a widely accepted framework for handling hyperintensional matters. Perhaps one ultimately needs "structured propositions," or "impossible worlds," or guise relativity, or something else entirely. But the best way that one can proceed at present with hyperintensional matters is to offer an account in the usual intensional idiom and trust that – should a framework for hyperintensionality emerge – the account will prove smoothly extendible. *Lewisian Knowledge* proceeds in exactly this best way.

It should be stressed that the intensional framework – as developed by Lewis and others – represents our best developed semantic framework to date. *It is no objection to an account of knowledge ascriptions that it is well integrated with our best developed semantic framework to date*. It would be an objection if it could be shown that the success of the account was somehow crucially reliant on known failings of our best framework, and so could not possibly extend to any potential successor frameworks. But nothing like that has been shown.

Indeed, suppose one chooses to extend the usual space of "possible worlds" by tacking on an outer sphere of "impossible worlds." Then the proposition expressed by '$5 + 7 = 12$' will differ from the proposition expressed by 'Fermat's Last Theorem holds', insofar as there will be mathematically impossible worlds in which the former holds but not the latter (and worlds in which the latter holds but not the former, and worlds in which neither holds). Should such a framework emerge as the best framework for hyperintensionality, then relevant alternatives theory would extend perfectly smoothly. I don't mean to suggest that this is the best framework for hyperintensionality (the jury is still out), or that relevant alternatives theory requires this to be the best framework for hyperintensionality (compatibility with frameworks has to be assessed case-by-case). But I take this example to show that relevant alternatives theory can extend very smoothly to at least some potential successor frameworks. And so I would conclude that, with respect to hyperintensionality, relevant alternatives theory is actually doing *as well as is currently possible*.

30.2.2 Truth Conditions

Even given a framework of possibilities, one might wonder whether the truth conditions for a knowledge ascription can be *equated* with the condition of the subject's evidence eliminating the relevant

alternatives, as per *Lewisian relevant alternatives theory*. There are a number of reasons that might be given for doubting the equation, but I will focus on the problem of *cheap knowledge*, arising in contexts in which the relevant alternatives are eliminated too easily.[13] In some cases of cheap knowledge the only relevant alternatives are *silly*. For instance, if Ann has never even heard of either David Lewis or of Princeton University, it might thereby seem false for anyone to say:

(4) Ann knows that David Lewis was a professor at Princeton University.

But suppose that Ann is a rabid baseball fan and has full knowledge of everyone who has ever played shortstop for the Chicago Cubs. Then in a context *c* in which David Lewis's being a shortstop for the Chicago Cubs is the only relevant alternative to him being a professor at Princeton University, *Lewisian relevant alternatives theory* has it that (4) is true in *c*.[14] In other cases of cheap knowledge the relevant alternatives are *empty* (either because there are no $\sim p$-worlds, or because none happen to be relevant). For instance, in a context *c* in which no alternatives to David Lewis being a professor at Princeton University are relevant, *Lewisian relevant alternatives theory* again has it that (4) is true in *c*, without requiring Ann to have any evidence whatsoever. Where the relevant alternatives are empty everyone trivially satisfies the truth conditions given by *Lewisian relevant alternatives theory*.

One tempting idea for solving the problem of cheap knowledge is to tinker with *Lewisian contextualism*, so as to ensure that there are always substantive relevant alternatives. But given that there are any propositions that are true at all worlds and that require substantive evidence to be known (e.g. that Fermat's Last Theorem holds), this move is hopeless. Given that the proposition in question is true at all worlds, no amount of tinkering with the rules of relevance can make an alternative relevant, simply because there are no alternatives out there that might be made relevant.[15]

Let me suggest an alternative move on Lewis's behalf, which involves tinkering with *Lewisian relevant alternative theory* to add further requirements (in ways that admittedly sacrifice some of its elegance, and reintroduce some of the epistemic requirements that Lewis sought to banish). For instance, one might reintroduce a belief requirement, as well as a basing requirement, as per:

> **Lewisian relevant alternatives theory, modified:** A sentence of the form '*s* knows that *p*' is true in context *c* iff (i) *s*'s evidence eliminates every not-*p* possibility relevant in *c*, and (ii) *s* believes that *p* on this basis.

For instance, if Ann has never heard of David Lewis or Princeton University, then even though Ann's evidence might eliminate every possibility in which David Lewis was not a professor at Princeton University in a given context *c* (either because *c* accords no relevant alternatives to this claim or silly alternatives), (4) will still come out false in *c*, simply because Ann will lack the relevant belief, or in more complicated cases will lack the right basis for this belief.[16]

Lewisian relevant alternatives theory, modified still allows that a subject can have *cheap evidence*. For instance, Ann is still accorded all the evidence required by (4) in certain contexts. This might seem itself problematic, though I think that the notion of evidence in play is partially a term of art. What *Lewisian relevant alternatives theory, modified* does is to prevent cheap evidence from automatically transmuting into cheap knowledge, simply by requiring more than evidence for knowledge.

30.2.3 Knowledge Itself

Even given a framework of possibilities and the equation of the truth conditions for a knowledge ascription with the condition of the subject's evidence eliminating the relevant alternatives, there remains a question of *knowledge itself*. When '*s* knows that *p*' is true in *c*, what – metaphysically

speaking – must hold in the world? (This question is obscured by Lewis's simplifying device of speaking as if he were discussing knowledge itself: section 30.1.1. One must first see that Lewis is only discussing knowledge ascription, in order to see that Lewis has left open the status of knowledge itself.)

There are at least three main sorts of answers that Lewis might give concerning the metaphysics of knowledge. One answer – that of the *epistemic pluralist* – begins by positing a plurality of binary knowledge-like relations K_1–K_n. What varies with context is then which epistemic relation K_j a given occurrence of 'know' denotes. This is to treat 'know' as having a contextually variable denotation, and the world as sporting a plurality of knowledge-like epistemic relations. But a second answer – that of the *epistemic monist* – posits a single ternary knowledge relation *Kspx*. What varies with context is which value for the third relatum *x* is expressed by an '*s* knows that *p*' sentence at a given context. This is to treat 'know' as having an invariant denotation, and the world as sporting a single knowledge relation, albeit one that is more complex than usually thought. And a third answer – that of the *epistemic nihilist* – posits no knowledge-like relations whatsoever. This theory posits only information about which alternatives various evidence would eliminate. What varies with context is the connection between an '*s* knows that *p*' sentence and the eliminative information it conveys. So in terms of the metaphysics of the knowledge relation, is Lewis a pluralist, monist, or nihilist about knowledge itself?[17]

I don't believe that anything in Lewis's discussion turns very heavily on the choice between pluralism, monism, and nihilism about knowledge itself. But it does seem to me that epistemic monism is independently the most plausible choice. Semantically, it fits the plausible idea that 'know' has an invariant denotation (this will be discussed further in section 30.3.2). And metaphysically, it fits the natural idea that there is a real unity to knowledge. In this respect the question of knowledge itself is merely a respect in which Lewis's own account is incomplete, not a respect in which it is problematic. Indeed I am suggesting that this gap in Lewis's account can be filled in a plausible way.

30.3 Semantic Questions

I have just considered three questions which might be thought of as epistemic questions, insofar as they primarily concern *Lewisian relevant alternatives theory*. I will conclude by considering four further questions for Lewis's account, which might be thought of as *semantic questions*, insofar as they primarily concern *Lewisian contextualism*. (I would reiterate that these questions are just some questions that any friend of contextualist relevant alternatives theory should address, and are not intended to exhaust every possible concern.)

30.3.1 'Know' Itself

Just as Lewis's account falls silent on knowledge itself, it likewise falls silent on the semantics of 'know' itself. *Lewisian knowledge ascription* offers truth conditions for a complex expression, of the form '*s* knows that *p*'. But presumably – given a compositional approach to semantics – the truth conditions for complex expressions are to be derived from the meanings of their component parts. Moreover, not all knowledge ascriptions have the form '*s* knows that *p*'. For instance, there are also knowledge-*wh* constructions such as:

(5) Ann knows what kind of bird is in the garden.

Lewisian knowledge ascription falls silent on these sentences. So what does Lewis think is the denotation of 'know' itself, which operates in all knowledge ascriptions (including sentences like 5), and from which *Lewisian knowledge ascription* should derive?

Lewisian contextualism describes the determinants of relevance, but leaves open how this connects to the denotation of 'know' itself. That said, there is a very natural treatment of the denotation of 'know' available to Lewis, which preserves the epistemic monist's idea of an invariant denotation (section 30.2.3), while extending his contextualism to other knowledge ascribing sentences, including (5). That is:

> $[[\text{know}]]^{c,w,t} = \lambda s\ \lambda p.$ *s*'s evidence eliminates every possibility in which not-*p* that is relevant in *c*.

On such a treatment, 'know' (relative to a context of utterance, and a world and time of evaluation[18]) invariably denotes a ternary relation between a subject, a proposition, and the contextually given relevant alternatives to the proposition (the contrasts), obtaining whenever the subject's evidence eliminates these contextually given relevant alternatives.

Or if one preferred to operate with *Lewisian relevant alternative theory, modified* (adding in belief and basing requirements), then one would reach:

> $[[\text{know}]]^{c,w,t} = \lambda s\ \lambda p.$ *s*'s evidence eliminates every possibility in which not-*p* that is relevant in *c*, and *s* believes that *p* on this basis.

Equivalently, one could treat the relevant possibilities in which not-*p* as a contrast proposition *q*:

> $[[\text{know}]]^{c,w,t} = \lambda s\ \lambda p.$ *s*'s evidence eliminates *q* (where *q* disjoins the possibilities in which not-*p* that are relevant in *c*), and *s* believes that *p* on this basis.

The result would be a contrastivist semantics for 'know' itself, in the spirit of Schaffer (2004b, 2005, 2007), and as developed in Schaffer & Szabó (forthcoming).[19]

So the matter of the semantics of 'know' itself is a further matter on which Lewis's account is incomplete but not necessarily problematic. Indeed I am suggesting that this gap in Lewis's account can be also filled in a plausible way.

30.3.2 Is 'Know' an Indexical?

Even given a contextualist denotation for 'know' itself (as suggested in the previous section), one might worry whether contextualism is generally semantically plausible. Does such a denotation look like anything else we find in semantics? Or is it an *ad hoc* invention? Lewis offers a range of guiding analogies for his contextualist proposal. He (1996, 563–4) speaks of 'know' as being like an indexical pronoun, like a gradable adjective, and – perhaps most centrally for his semantics – as being like a quantificational determiner with a contextually restricted domain argument. But these guiding analogies are inequivalent, and moreover each analogy is individually questionable.[20]

These guiding analogies are inequivalent, insofar as the leading treatments of indexical pronouns, gradable adjective, and quantificational determiners involve quite different structures. Indexical pronouns have no extra structure, but instead might be thought of (following Kaplan 1989) as having a context invariant character which is a function from contexts to contents. Gradable adjectives (at least on the treatment of Kennedy and McNally 2005) have additional structure in the form of a

degree argument, and associate with comparative and degree morphology, while quantificational determiners (on a view which traces back to Heim 1982) have quite different additional structure in the form of a restrictor argument.[21]

Moreover these guiding analogies are each individually questionable. Semantically, 'know' does not seem to behave like indexical pronouns such as 'I', 'here', and 'now', which are all obviously context sensitive in a way that 'know' is not, and which exhibit smooth tracking across contexts (e.g. we have no trouble tracking the referent of 'I' as speakers take turns in a conversation) in a way that 'know' does not. And syntactically 'know' does not seem to project either a degree argument or a restrictor argument, which (respectively) explain the context sensitivity of the semantics for gradable adjectives and quantificational determiners.

The resulting concern about the semantic plausibility of contextualism constitutes, to my mind, one of the main concerns with *Lewisian contextualism* (and contextualism generally). The friend of *Lewisian contextualism* might seek to defend one of the guiding analogies (for instance, see Blome-Tillmann 2008 and Ichikawa 2011). In my opinion the prospects for such a defense are grim, though I should acknowledge that this calls for further discussion. Or she might abandon the search for a guiding analogy, on grounds that our understanding of the context sensitivity of language is too primitive, and/or that the context sensitivity of 'know' might be *sui generis* (DeRose 2009, 13). But what is that, other than an admission that contextualism still looks – at least to the best of our limited current knowledge – to be implausible?

So I would suggest a different move on Lewis's behalf, which is to find a better guiding analogy. Along these lines, Schaffer and Szabó (forthcoming) suggest that 'know' is best understood as an *A-quantifier*, where A-quantifiers traditionally include *a*dverbial quantifiers, modal *a*uxiliaries, and – on the classic views of Hintikka 1969 and Heim 1992 – *a*ttitude verbs. Rather than having a syntactically projected restrictor argument as per quantificational determiners (D-quantifiers), A-quantifiers pick up their restrictors at the semantic level, via a "modal base" argument interpreted via general discourse features (the question under discussion: section 30.3.3). A-quantifiers display a characteristic range of features, including: sensitivity to the question under discussion, association with focus, ability to take explicit 'if'-clause as restrictors, and ability to coordinate domains with other A-quantifiers. Schaffer and Szabó then argue that 'know' displays these characteristic marks of being an A-quantifier, though it would only be fair to say that "the jury is still out" on whether an A-quantification model – or any other model – will prove viable.

30.3.3 Deriving Relevance

Even given a contextualist denotation for 'know' itself, augmented with a plausible guiding analogy, there remains room to question the specific list of rules that comprise *Lewisian contextualism*. One might question the details, by adding, subtracting, and/or revising a handful of the rules on the list Lewis provides.[22] But at a deeper level – which I will focus on – one might question the entire project. For it seems that what Lewis is giving us in the end, however exactly the details are settled, is a somewhat ad hoc laundry list of rules specifically tailored to the epistemic case (Ichikawa 2011, 386–7). One might think – given that relevance plays systematic rules throughout semantics – that a theory of relevant alternatives in epistemology should be *derived* from a single general notion of relevance at work elsewhere in semantics.

Lewis sometimes speaks as if his rules of relevance should be derived from the context sensitivity of 'every'. In this vein, Lewis (1996, 553) says that his requirement that "every" not-*p* possibility be eliminated is the source of his contextualism: "[W]e must attend to the word 'every.' What does it mean to say that every possibility in which not-*P* is eliminated? An idiom of quantification, like 'every,' is normally restricted to some limited domain." But Lewis in fact makes no attempt to show

that any of his specific rules derive from the general context sensitivity involved with 'every', nor could he since many of his rules are tailored to the epistemic case.

Indeed – as Stanley (2005, 66) has argued – among the general features of the context sensitivity of quantificational determiners like 'every' is a natural capacity to shift freely between occurrences. Thus one finds sentences such as:

(6) Every sailor waved to every sailor.

There is a natural reading of (6) in which the first domain of sailors and the second domain of sailors are entirely disjoint, for instance if the speaker is discussing a scenario in which two ships passed at sea and every sailor on the first ship waved to every sailor on the second ship. But among the specific features of the context sensitivity of 'know' that Lewis posits is a lack of such free shifting. Specifically, once skeptical scenarios come into play, subsequent knowledge ascriptions in that discussion are supposed to go false (as befitting what Lewis (1996, 551) calls the "irresistible" force of skepticism), which – given *Lewisian knowledge ascription* – requires the skeptical scenarios to remain fixedly relevant for as long as the discussion persists. Indeed Lewis (1979, 355) speaks of this in terms of a general (and somewhat mysterious) asymmetry in accommodation: "[F]or some reason raising of standards goes more smoothly than lowering." And: "Because of this asymmetry, a player of language games who is so inclined may get away with it if he tries to raise the standards of precision as high as possible…". So it seems as if the context dependence that Lewis himself posits for 'know' works very differently from the context dependence one finds for quantificational determiners like 'every'.

I would suggest – as a way to derive the epistemically relevant alternatives from a single general notion of relevance at work elsewhere in semantics, while explaining the lack of free shifting – reading the epistemically relevant alternatives off *the question under discussion*. That is, I would suggest:

Lewisian contextualism, modified: A possibility is relevant at a context if it is a possible answer to the question under discussion in that context.

The question under discussion is a fairly orthodox semantic posit, posited as an entry on Lewis's (1979) "conversational scoreboard" to reflect what is being addressed at that point in the conversation. This posit serves to explain diverse phenomena including felicitous topicalization, distant ellipsis, Gricean relevance ("speak to the question"), relevant alternative sets for contrastive focus, and domain restriction for A-quantification (Carlson 1983; Ginzburg 1996; Roberts 2004; see Schaffer and Szabó forthcoming: section 30.3.3 for further applications). *Lewisian contextualism, modified* would thus derive the epistemically relevant alternatives in way that fits a general notion of relevance at work elsewhere in semantics. Indeed this thesis is especially fitting for a view which models 'know' as an A-quantifier, as per the Schaffer and Szabó view (§3.2).

Lewisian contextualism, modified would also explain the lack of free shifting, since the question under discussion is a relatively stable discourse-level matter (in contrast with domain restrictors for D-quantifiers like 'every', which presumably have independent syntactic realizations that can freely be coordinated or not). Indeed there is independent evidence that A-quantifiers like modals do not permit free shifting (Stanley 2005, 73), as seen in the unacceptability of sentences such as:

(7) Ann can speak Finnish and Ann can only speak English.

If the question under discussion could freely shift from *whether Ann has the capacity to learn Finnish if she studies for years*, to *whether Ann in fact has learned any Finnish*, then (7) should be acceptable. The unacceptability of (7) thus provides evidence against free shift for the question under discussion.

From a deeper perspective, *Lewisian contextualism, modified* fits a conception of our concept of knowledge as used to indicate *who can answer the question*. Recall:

(1) Ann knows that there is a goldfinch in the garden.

Whether (1) is true in a given context seems to depend on whether the question under discussion is the easy question of *whether there is a goldfinch or a raven in the garden*, or the hard question of *whether there is a goldfinch or a canary in the garden*. This fits the intuitive shiftiness of knowledge ascriptions (section 30.1.3), Schaffer's (2005, §1) conception of the role of knowledge as *fingering answerers*, and Hookway's (1996, 7) insight into the general role that knowledge plays in *evaluating inquiry*: "The central focus of epistemic evaluation is . . . the activity of inquiry . . . When we conduct an inquiry . . . we attempt to formulate questions and to answer them correctly."

Lewisian contextualism, modified, however, does not support Lewis's *Belief* or *Actuality* rules, since these considerations are not in general required to be factored into possible answers to the question under discussion. As a result, *Lewisian contextualism, modified* cannot sustain Lewis's idea of shunting the justification, truth, and belief conditions into considerations of relevance (section 30.1.2), and so would require something like *Lewisian relevant alternatives theory, modified* instead (section 30.2.2). Indeed it would require a further modification that included a truth condition, such as:

> **Lewisian relevant alternatives theory, re-modified:** A sentence of the form '*s* knows that *p*' is true in context *c* iff (i) *s*'s evidence eliminates every not-*p* possibility relevant in *c*, (ii) *s* believes that *p* on this basis, and (iii) *p* is true.

In this respect it seems to me that *Lewisian contextualism, modified* is generating useful constraints on the shape of the accompanying relevant alternatives theory, and can thereby help the friend of contextualist relevant alternatives theory settle the question of whether to treat a given factor via the contextualist aspect of her theory, or via the relevant alternatives aspect of her theory.

30.3.4 How Did He Do It?

I began my statement of Lewis's account (section 30.1.1) by noting that he states his account in terms that he later reveals to be an incorrect simplification, in speaking of when *s* knows that *p*, instead of the contexts in which '*s* knows that *p*' counts as true. Though Lewis presents this as something of a harmless simplification (which the careful reader can correct if she likes), I think that Lewis's simplified statement actually raises a serious and underappreciated problem for him, concerning how it could possibly have gone well (by Lewis's own lights).

The problem is that Lewis is presenting his theory in a philosophical context in which skeptical scenarios are explicitly under attention and therefore relevant (by his irresistible rule of *Attention*).[23] By *Lewisian knowledge ascription* (plus some plausible assumptions about the uneliminability of such skeptical scenarios, which Lewis himself embraces) this entails that virtually all knowledge ascriptions made in Lewis's paper are false. Yet Lewis's simplified object language exposition involves various positive claims of knowledge. So doesn't the theory itself predict that the simplified

presentation Lewis opts for should fail? That is, doesn't the theory itself predict that the reader – or at least the accommodating reader, who allows Lewis to make skeptical scenarios salient – should be plunged into a skeptical context and thereby deem most if not all knowledge claims to be false? And doesn't this conflict with the fact that Lewis makes positive knowledge claims in such a context, and that many if not all of these positive knowledge claims still seem true in the skeptical context Lewis has created?

Lewis (1996, 566–7) sees the problem: "Don't you smell a rat? Haven't I, by my own lights, been saying what cannot be said?" and "Does not my story deconstruct itself?" His (1996, 566) reply is that he has "bent the rules," but was able to do so by relying on "the cardinal rule of pragmatics, which overrides every one of the rules I mentioned: interpret the message to make it make sense – to make it consistent, and sensible to say." (This is the portion at the close of the paper at which Lewis mentions that he has offered a simplified and potentially misleading presentation; it is also the portion – alluded to in section 30.1.2 – at which Lewis renders all of his rules of relevance resistible.)

But Lewis's reply – and the general idea that one can bend the rules in this way via "the cardinal rule of pragmatics" seems to conflict with certain of Lewis's key tenets, especially Lewis's (1996, 561) motivating puzzle: "how can it be, when his conclusion is so silly, that the sceptic's argument is so irresistible?" For now Lewis's "cardinal rule" is making it possible to get away with ordinary knowledge ascriptions in a context in which skeptical doubts are relevant (indeed that is exactly what Lewis is getting away with). Thus this response seems to license what DeRose (1995, 28) labels "abominable conjunctions" such as "I don't know that I'm not a bodiless (and therefore handless) brain-in-a-vat, but I still know that I have hands." Such abominable conjunctions should seem felicitous enough if one can just "bend the rules" when one reaches the second conjunct.[24]

It seems to me that – by the lights of most plausible contextualisms, or at least those that are skeptic-friendly in allowing the skeptic to succeed in the contexts she creates – Lewis's presentation simply ought not to have succeeded. The theory predicts that the reader will judge Lewis's positive knowledge claims to be false, and so regard his discussion as an outright failure.

One option for the contextualist is to be less skeptic-friendly, and allow ordinary knowledge ascriptions to still count as true even when skeptical scenarios are relevant. This option certainly counts as a departure from Lewis's approach. And worse, on this option the contextualist may need to abandon her claim to reconcile ordinary knowledge with skeptical doubt, and may need to withdraw her claim to meet Lewis's (1996, 550) primary motivation of steering between "the rock of fallibilism and the whirlpool of skepticism."[25]

A second option for the contextualist is to chalk this up to *compartmentalization*. In this vein Lewis (1996, 565) imagines two epistemologists on a bushwalk, whose discussion of skepticism might be interspersed with claims that they know where they are going, or that they know which sort of viper slithers yonder, with little to no "relevance leakage" between these discussions. I think that compartmentalization is plausible for this sort of case, but would only add that Lewis's discussion in "Elusive Knowledge" is a single extended and unified discussion (uninterrupted by the appearance of lethal wildlife or other conversation-stoppers) for which compartmentalization is inappropriate. Indeed, if the contextualist allows for compartmentalization to occur easily within a single unified discussion, she will be in danger of re-licensing abominable conjunctions so long as one can just "switch compartments" between the conjuncts.

Though I am sympathetic with the broad outlines of Lewis's contextualist relevant alternatives theory, I find it hard to avoid thinking that the success of Lewis's own simplified presentation is actually a deep embarrassment for his contextualism, one for which an explanation remains elusive.

Acknowledgments

Thanks to Michael Blome-Tillmann, Stew Cohen, Keith DeRose, Jonathan Ichikawa, and Rae Langton.

Notes

1 For instance, both Langton (2004) and Schaffer (2004c) argue that Lewis's contextualism does not fit his skeptical worries about properties in "Ramseyan Humility," since his contextualism should dissolve these skeptical worries.

2 As Lewis (1973, 92) clarifies, in reply to Quinean concerns that counterfactuals are too shifty for serious treatment: "I am not one of those philosophers who seek to rest fixed distinctions upon a foundation quite incapable of supporting them. I rather seek to rest an unfixed distinction upon a swaying foundation, claiming that the two sway together rather than independently."

3 One might think that these topics are so closely connected as to never differ. For one might think that, by disquotation, '*s* knows that *p*' is true in *c* iff *s* knows that *p*. But disquotation so formulated fails for contextually sensitive terms (for instance, ['I am Schaffer' is true iff I am Schaffer] can be false if I am the speaker and you are the disquoter). And by Lewis's contextualist lights, 'know' is a contextually sensitive term.

4 In this respect, Lewis's account differs from most other classic relevant alternatives theoretic accounts (cf. Austin 1946; Goldman 1976; and Dretske 1981), which really are object language accounts of when *s* knows that *p*.

5 By way of comparison, Goldman (1976, 772) starts from the idea that "a person is said to know that *p* just in case he *distinguishes* or *discriminates* the truth of *p* from relevant alternatives." But he (1976, 785–6) ultimately embeds this condition in clauses (3.c.i) and (3.c.ii) of a vastly more complicated account, and one which is moreover restricted to the special case of non-inferential perceptual knowledge of an object having a property.

6 The list of expanders (/contractors) is the same as the list of irresistibles (/resistibles). This seems largely accidental. As far as I can tell, Lewis has no deep theoretical reasons for rejecting resistible expanders, and no deep reason for rejecting irresistible contractors, so long as these can be insulated from conflict with whatever irresistible expanders there might be.

7 For instance, Lewis leaves some of the details of *Resemblance* open. As Lewis notes, *Resemblance* is in danger of overgenerating relevant possibilities in a way that would lead to rampant skepticism. After all, if resemblance with respect to the subject's evidence is salient, then – at least on an "internalist" conception of evidence on which the subject has the same evidence when envatted as when embodied (a conception Lewis endorses on p. 556) – skeptical scenarios will be relevant whenever the subject's evidence is salient. So Lewis concludes that *Resemblance* needs some yet-to-be-specified restriction.

8 The full story: *Actuality* ensures that the subject *s*'s world is relevant. Lewis's conception of elimination (1996, 553) entails that *s*'s evidence can never eliminate her actuality. Together these ensure that if *s* is at a not-*p* world, then *s*'s world will be relevant (by *Actuality*) and uneliminated (by the conception of elimination), which then suffices (by *Lewisian Knowledge*) for '*s* knows that *p*' to be false in all contexts.

9 Against a justification requirement, Lewis (1996, 551) asks: "What (non-circular) argument supports our reliance on perception, on memory, and on testimony?" He also talks about knowledge supported by forgotten reasons, and adds: "[W]e know the name that goes with the face, or the sex of the chicken, by relying on subtle visual cues, without knowing what those cues may be." Against a belief requirement, Lewis (1996, 556) invokes cases like that of Radford's (1966) timid student, who gives the right answer but feels no confidence in what she says.

10 Schaffer (2004a) plays with this idea on behalf of the skeptic, with the idea being that knowledge ascriptions, though virtually all *false*, still may be *felicitous* in ordinary contexts on the model of felicitous hyperbole. On this treatment Lewis's rules of relevance are re-understood as felicity conditions on hyperbole.

11 In this vein, Schaffer and Knobe 2012 (replicated by Buckwalter 2014) present a range of experimental results confirming a strong, stable, and unified pattern of intuitive sensitivity to the contextually relevant options.
12 For pragmatic accounts, see Rysiew 2001 and Brown 2005. For performance error accounts, see Nagel 2008 and Bach 2010.
13 See Lihoreau (2008) for a version of this objection to Lewis.
14 Though it may be that Lewis's rule of *Relevance* will preclude there being such a context, by forcing some further relevant alternatives, such as Lewis being a professor at Harvard University (I thank Jonathan Ichikawa for discussion on this point).
15 Though on certain ways of handling hyperintensionality, such as by having an indefinitely extensible sphere of impossible worlds, every proposition can be accorded alternatives. So there is room to try to handle hyperintensionality and resolve the problem of cheap knowledge by (i) having an indefinitely extensible sphere of impossible worlds which generates alternatives for any proposition, and (ii) tinkering with the rules of relevance to ensure that certain alternatives are always relevant, e.g. by having an irresistible expander that renders the nearest alternative(s) relevant: ("The nearest not-*p* possibility(s) are always relevant.")
16 Additional motivation for inserting a belief and basing requirement comes from cases of *missed clues* (Schaffer 2001), in which the subject's evidence indeed eliminates the relevant alternatives, but the subject herself has no appreciation of this fact.
17 One can find strands in Lewis's thought that might seem to favor each of these three perspectives. Epistemic pluralism fits Lewis's (1986, 59–60) general picture of the abundance of (non-fundamental) properties and relations, there being "one of them for any condition we could write down, . . . ". Epistemic monism fits Lewis's (pers. comm.; cf. Schaffer 2004b, 97) embrace of my contrastivism, which I had proposed as a monistic rival: "The only thing we disagree about is whether we disagree." And epistemic nihilism fits Lewis's (1996, 563) concluding discussion of "What is it all for?" where he says: "Ascriptions of knowledge . . . are a very sloppy way of conveying very incomplete information about the elimination of possibilities."
18 I insert world and time indices as per Kaplan's orthodox semantic framework. In fact Lewis (1980) preferred to operate with not just world and time indices, but additional indices for location and standard-of-precision.
19 For related versions of contrastivism, see Morton and Karjalainen 2003 and Sinnott-Armstrong 2004.
20 The worry about semantic plausibility represents perhaps the leading objection to contextualism in the current literature. Thus Stanley (2005, 47) – one of the main proponents of this objection – maintains: "[T]he alleged context-sensitivity of knowledge ascriptions has no other parallel among the class of uncontroversial context-sensitive expressions." And Blome-Tillmann (2008, 29) – though himself defending the idea that 'know' is an indexical – acknowledges: "Epistemic Contextualism . . . has fallen into considerable disrepute recently. Many theorists have raised doubts as to whether 'know' is context-sensitive, typically basing their arguments on data suggesting that 'know' behaves semantically and syntactically in a way quite different from recognized indexicals . . . ". See Schaffer and Szabó (forthcoming, esp. §2) for further discussion.
21 Lewis (1996, 554) thinks of gradable adjectives in quantificational terms, via: "Just as *P* is known iff there are no uneliminated possibilities of error, so likewise a surface is flat iff there are no bumps on it. We must add the proviso: Psst! – except for those bumps that we are properly ignoring." But on leading views (Kennedy and McNally 2005) gradable adjectives are not quantifiers but degree-relative notions. What makes a surface flat is not the number of bumps it has, but rather its degree of bumpiness.
22 Along these lines, Cohen 1998 and Blome-Tillmann 2009 each propose tweaks to Lewis's rules.
23 Indeed these scenarios come up for discussion on the very first page of "Elusive Knowledge," when Lewis (1996, 549) instructs us to let our "paranoid fantasies rip – CIA plots, hallucinogens in the tap water, conspiracies to deceive, old Nick himself . . . ". Discussion of such scenarios recurs throughout Lewis's discussion.
24 Stanley (2005, 67) argues that Lewis's theory licenses abominable conjunctions, on grounds that (i) Lewis treats 'know' as an indexical (cf. section 30.3.2), and (ii) indexicals are capable of freely shifting within a discourse. Stanley takes this to show that Lewis should not treat 'know' as an indexical. What I'm pointing

out is an entirely separate reason why Lewis is committed to abominable conjunctions, which Lewis could not block simply by rejecting the treatment of 'know' as an indexical.

25 Though skeptic-unfriendly versions of contextualism might still be said to reconcile ordinary knowledge with skeptical doubt, insofar as they can still be said to explain the mechanism by which the skeptic's claims can count as true, even if that mechanism turns out to be seldom activated or to be often thwarted by other conversational pressures. Thus DeRose (1995, 6) says: "[T]he important point is to identify the mechanism by which the skeptic at least threatens to raise the standards for knowledge. Whether the skeptic actually succeeds against a determined opponent in so raising the standards is of little importance." See DeRose (2009, ch. 4) for a more detailed discussion of options for the contextualist, including DeRose's preferred "Gap view" (2009, 144–8), on which the skeptic and her opponent thwart each other from reaching conversational equilibrium, so that neither speaks truly (or falsely either). I am grateful to Keith DeRose for discussion on this point.

References

Austin, J.L. (1946). Other minds. *Proceedings of the Aristotelian Society*, 20, 149–87.

Bach, K. (2010). Knowledge in and out of context. In J.K. Campbell, M. O'Rourke & H. Silverstein (Eds.), *Knowledge and skepticism*, pp. 105–36. Cambridge, MA: MIT Press.

Blome-Tillmann, M. (2008). The indexicality of 'knowledge'. *Philosophical Studies*, 138, 29–53.

Blome-Tillmann, M. (2009). Knowledge and presuppositions. *Mind*, 118, 241–94.

Brown, J. (2005). Comparing contextualism and invariantism on the correctness of the contextualist intuitions. *Grazer Philosophische Studien*, 69, 71–100.

Buckwalter, W. (2014). The mystery of stakes and error in ascriber intuitions. In J. Beebe (Ed.), *Advances in experimental epistemology*, pp. 145–74. New York: Bloomsbury.

Carlson, L. (1983). *Dialogue games: an approach to discourse analysis*. Dordrecht: Reidel.

Cohen, S. (1988). How to be a fallibilist. *Philosophical Perspectives*, 2, 91–123.

Cohen, S. (1998). Contextualist solutions to epistemological problems: scepticism, Gettier, and the lottery. *Australasian Journal of Philosophy*, 76, 289–306.

DeRose, K. (1995). Solving the skeptical problem. *Philosophical Review*, 104, 1–52.

DeRose, K. (2009). *The case for contextualism: knowledge, skepticism, and context*, Vol. 1. Oxford: Oxford University Press.

Dretske, F. (1981). The pragmatic dimension of knowledge. *Philosophical Studies*, 40, 363–78.

Ginzburg, J. (1996). The semantics of interrogatives. In S. Lappin (Ed.), *The handbook of contemporary semantic theory*, pp. 385–422. Oxford: Blackwell.

Goldman, A. (1976). Discrimination and perceptual knowledge. *Journal of Philosophy*, 73, 771–91.

Hawthorne, J. (2004). *Knowledge and lotteries*. Oxford: Oxford University Press.

Heim, I. (1982). The semantics of definite and indefinite noun phrases. PhD dissertation, University of Massachusetts-Amherst.

Heim, I. (1992). Presupposition projection and the semantics of attitude verbs. *Journal of Semantics*, 9, 183–221.

Hintikka, J. (1969). Semantics for propositional attitudes. In J.W. Davis, D.J. Hockney & W.K. Wilson (Eds.), *Philosophical logic*, pp. 21–45. Dordrecht: Reidel.

Hookway, C. (1996). Questions of context. *Proceedings of the Aristotelian Society*, 96, 1–16.

Ichikawa, J. (2011). Quantifiers and epistemic contextualism. *Philosophical Studies*, 155, 383–98.

Kaplan, D. (1989). Demonstratives, an essay on the semantics, logic, metaphysics, and epistemology of demonstratives and other indexicals. In J. Almog, J. Perry & H. Wettstein (Eds.), *Themes from Kaplan*, pp. 481–563. Oxford: Oxford University Press.

Kennedy, C. & McNally, L. (2005). Scale structure and the semantic typology of gradable predicates. *Language*, 81, 345–81.

Langton, R. (2004). Elusive knowledge of things in themselves. In F. Jackson & G. Priest (Eds.), *Lewisian themes: the philosophy of David K. Lewis*, pp. 130–7. Oxford: Oxford University Press.

Lewis, D. (1969). *Convention: a philosophical study*. Cambridge, MA: Harvard University Press.
Lewis, D. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic*, 8, 339–59.
Lewis, D. (1980). A subjectivist's guide to objective chance. In R. Jeffrey (Ed.), *Studies in inductive logic and probability*, Vol. 2, pp. 83–132. Berkeley: University of California Press.
Lewis, D. (1982). Whether report. In T. Pauli (Ed.), *Philosophical essays dedicated to Lennart Åqvist on his fiftieth birthday*, pp. 194–206. Uppsala: Filosofiska Studier.
Lewis, D. (1986). *On the plurality of worlds*. Oxford: Blackwell.
Lewis, D. (1988a). What experience teaches. *Proceedings of the Russellian Society*, 13, 29–57.
Lewis, D. (1988b). Desire as belief. *Mind*, 97, 323–32.
Lewis, D. (1996). Elusive knowledge. *Australasian Journal of Philosophy*, 74, 549–67.
Lewis, D. (1999). Why conditionalize? In D. Lewis, *Papers in metaphysics and epistemology*, 403–7. Cambridge: Cambridge University Press.
Lewis, D. (2001). Sleeping Beauty: reply to Elga. *Analysis*, 61, 171–6.
Lewis, D. (2009). Ramseyan humility. In D. Braddon-Mitchell & R. Nola (Eds.), *Conceptual analysis and philosophical naturalism*, pp. 203–22. Cambridge, MA: MIT Press.
Lihoreau, F. (2008). Relevant alternatives contextualism and ordinary contingent knowledge. *Disputatio*, 24, 281–94.
Morton, A. & Karjalainen, A. (2003). Contrastive knowledge. *Philosophical Explorations*, 6, 74–89.
Nagel, J. (2008). Knowledge ascriptions and the psychological consequences of changing stakes. *Australasian Journal of Philosophy*, 86, 279–94.
Radford, C. (1966). Knowledge – by examples. *Analysis*, 27, 1–11.
Roberts, C. (2004). Context in dynamic interpretation. In L. Horn & G. Ward (Eds.), *The Handbook of Pragmatics*, pp. 197–220. Oxford: Blackwell.
Rysiew, P. (2001). The context-sensitivity of knowledge attributions. *Noûs*, 35, 477–514.
Schaffer, J. (2001). Knowledge, relevant alternatives, and missed clues. *Analysis*, 61, 202–8.
Schaffer, J. (2004a). Skepticism, contextualism, and discrimination. *Philosophy and Phenomenological Research*, 69, 138–55.
Schaffer, J. (2004b). From contextualism to contrastivism. *Philosophical Studies*, 119, 73–103.
Schaffer, J. (2004c). Quiddistic knowledge. In F. Jackson & G. Priest (Eds.), *Lewisian themes: the philosophy of David K. Lewis*, pp. 210–30. Oxford: Oxford University Press.
Schaffer, J. (2005). Contrastive knowledge. *Oxford Studies in Epistemology*, 1, 235–71.
Schaffer, J. (2007). Knowing the answer. *Philosophy and Phenomenological Research*, 75, 383–403.
Schaffer, J. & Knobe, J. (2012). Contrastive knowledge surveyed. *Noûs*, 46, 675–708.
Schaffer, J. & Szabó, Z.G. (2014). Epistemic comparativism: a contextualist semantics for knowledge ascriptions. *Philosophical Studies*, 168, 491–543.
Sinnott-Armstrong, W. (2004). *Pyrrhonian skepticism*. Oxford: Oxford University Press.
Sosa, E. (1986). On knowledge and context. *Journal of Philosophy*, 83, 584–5.
Stanley, J. (2005). *Knowledge and practical interests*. Oxford: Oxford University Press.
Unger, P. (1975). *Ignorance: a case for scepticism*. Oxford: Oxford University Press.

31

Humility and Coexistence in Kant and Lewis

Two Modal Themes, with Variations

RAE LANGTON

Introduction

What could Immanuel Kant and David Lewis have in common? At first sight, little enough, given that Kant did more than most to chasten the ambitions of metaphysics, and Lewis did more than most to indulge them. Nevertheless I want to pursue here two common themes: first, an argument about ignorance of things in themselves, viewed as a kind of epistemic *humility*; and, second, an argument – perhaps even a transcendental argument – about the conditions of *coexistence*, the relation that worldmates bear to each other. The first commonality was flagged by Lewis himself, and has gone on to receive attention, while the other has gone so far unremarked.

Kant thinks we are ignorant of things "in themselves." This ignorance is not idealism or skepticism, but it does concern limits on our knowledge. That was my claim in *Kantian Humility*: according to Kant, much is known to us, but things "in themselves" – the *intrinsic* properties of things – remain hidden.[1] I argued that according to Kant, this ignorance, this Humility, arises from a contingency.[2] The idea, roughly, is that how things are intrinsically doesn't fix how things are relationally; and this contingency has Humility as its consequence, since knowledge of relational, or "phenomenal," facts fails to yield knowledge of intrinsic facts. Lewis agrees with Kant about our ignorance. Kant's claim that we can "find out nothing about [things] as they are in themselves" is "true . . . or at least something very like it is," he writes in "Ramseyan Humility," drawing an explicit parallel between Kant's conclusion and his own.[3] And for Lewis too our ignorance arises from a contingency.[4] The contingent connection between relational "role" properties and their intrinsic "realizers" means that our knowledge of the former fails to deliver knowledge of the latter. So for Lewis, as for Kant, a contingent connection between relational and intrinsic facts is seen to have epistemological consequences. This, then, is our first theme: Kant and Lewis are alike in their commitments, not only to Humility, but to Humility's source.

The second theme, that of necessity and coexistence, was an enduring topic of interest for Kant. In his early work, he explored the metaphysically necessary conditions of coexistence in his early work, and later, the "transcendentally" necessary conditions of our *experience* of coexistence. It is of

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

interest in this context, partly for its association with Kant's early argument for the aforementioned contingency; partly for its illustration of how metaphysical necessity resurfaces as transcendental necessity, in the transition from pre-critical to critical incarnations of Kant's thinking; but above all for providing the occasion for a further, surprising, commonality with Lewis. We shall see that it is not too far-fetched to interpret Lewis as offering an argument comparable to Kant's – perhaps even a transcendental argument – about the necessity of certain conditions for coexistence.[5]

What will emerge here is a Lewis who is more like Kant than you might have thought: a Lewis who is interested in the necessary, perhaps "transcendentally" necessary, conditions for coexistence; who denies knowledge of things in themselves; and who lays this ignorance at the door of a contingent connection between intrinsic and relational facts, somehow conceived. Needless to say, the differences between Kant and Lewis are plentiful and profound – now there's an understatement! – but I shall be skating over differences lightly, for the story's sake.

31.1 Contingency and Humility

31.1.1 Kant

Many of Kant's early writings explore a kind of contingency that troubled him long before Hume is supposed to have awoken him from his dogmatic slumbers. Kant's early dynamic theory of matter is developed against a backdrop of independent "simple" substances, descendants of Leibniz's monads, which possess two classes of properties: intrinsic properties, and "relational determinations," conceived of as physical forces, whose dynamic interactions constitute "phenomenal" substance or matter. These "relational determinations" of substance are regarded as "not necessary" to the substances as independent beings.[6] This contingency forms the centerpiece of Kant's early *New Exposition of the First Principles of Metaphysical Knowledge* (1986a [1755]):

> Single substances, of which neither is the cause of the existence of the other, have a separate existence, i.e. an existence that is absolutely intelligible without all the others. Therefore if the existence simpliciter of a substance is posited, there is nothing in that which proves the existence of other substances different from itself. Indeed, since relation is a determination that looks toward something else (i.e. it will not be intelligible in a being viewed entirely by itself), the relation and its determining reason cannot be understood through the existence of the substance as posited in itself. If, therefore, nothing more is added to this existence, there would be no relation among beings and clearly no mutual interaction. Therefore, in so far as single substances have an existence independent of other substances, there is no place for their mutual connection . . . , and it must be granted that their relation depends on a common cause, namely God. (413)

Kant is arguing that there is no necessary entailment from facts about intrinsic properties of substances to the "mutual relations" holding between them. The "relational determinations" of substances are not given by the substances on their own, endowed with the intrinsic properties they have in existing on their own. The "relational determinations" are not intrinsic, and cannot be understood through the existence, *simpliciter*, of substances. If nothing is "added" to the existence of substances there would be "no relation among beings" (413). If a substance has a power to relate to, or affect, other substances, that is not something it achieves in virtue of its intrinsic properties: as he puts it, "a substance never has the power through its own intrinsic properties to determine others different from itself" (415). He continues that in order for relations among substances to exist, something must be "added" through an act of God, and that this creative act is "obviously arbitrary" on God's part and "can be omitted or not omitted at His own pleasure" (414).

The argument poses a number of interpretive questions, which we can raise without letting them delay us unduly. What are the "relational determinations" supposed to be? The level of abstraction tempts one to think Kant has in mind relations, or relational properties, in general. But we also learn that he has in mind some *physical* relational properties: he mentions Newtonian attraction, and it seems plausible to include also the second fundamental force of his matter theory, namely repulsion or impenetrability. There is a question whether his interest is in the relation of, for example, one thing *attracting* another; or in the relational, dispositional property, of, for example, one thing having the *power* to attract another.

A second question is whether the argument is about concepts, or about robust properties. Kant's talk of "relation" not being "intelligible" in a being "viewed entirely by itself" might suggest concepts, but then the talk of God's creation suggests robust metaphysics: in creating substance with its intrinsic properties, God has not done enough to create relations or relational properties.[7]

A third question: supposing that metaphysically robust properties are at stake, in what sense do the "relational determinations" – for example, "attraction" and "impenetrability" – fail to be intrinsic? While *attracting* is extrinsic, perhaps having a *power to attract* is not. The power appears to pass standard tests for intrinsicness, such as the isolation test: something existing all on its own can have the power to attract, whether or not it exercises that power on another coexisting thing. Yet Kant is at least sometimes talking about the power as being relational, and as requiring something more (415). It may be worth bringing an amended notion of intrinsicness to our interpretation, which would count such powers as extrinsic: perhaps an intrinsic property is one that can be possessed independent of other entities and *of laws* God might create.[8]

Fourth, there is a distributive/collective ambiguity. Does Kant mean that one substance, taken individually, is insufficient to establish that substance's "relational determinations"? (That's what I have elsewhere called "unilateral reducibility.") Or does he mean that all the substances, taken collectively, are insufficient to establish the "relational determinations" that hold among them? (That's what I have elsewhere called "bilateral reducibility.") There are textual reasons in favor of each of these, but they compete.

Many interpretive possibilities present themselves, but we need to narrow them down. So let us pursue the following thought on Kant's behalf, without further defense. He is proposing, at least, that the causal powers of attraction and impenetrability are not themselves intrinsic properties; and they are not necessitated by the intrinsic properties of substances, taken collectively. No matter how many substances God creates, with their intrinsic properties, that's not enough to establish their causal powers of attraction and impenetrability, and therefore not enough to establish physical relations among them. Note that it is compatible with this interpretation that certain relations and relational properties, such as similarity, are reducible to, or supervene on, the intrinsic properties of substances, and would not require a special act of creation.

While the contingency described here concerns causality, it is not the problem (or not the main problem) raised by Hume: it is not about the causally necessary connection, or absence thereof, between two events; rather it is about a metaphysically necessary connection, or absence thereof, between two classes of properties: causal powers and intrinsic properties.[9] Nor does this contingency concern the modality of propositions, first and foremost, but rather of properties, features of the world. Talking about what God could or couldn't create provides a natural vehicle for its discussion.

In affirming this contingency Kant contradicts Leibniz, who upheld a necessary connection between physical phenomena and monadic things in themselves, denying their "distinct existence" (to borrow Hume's phrase). For Leibniz the realm of physical phenomena coincides with the realm of monadic things in themselves – as Kant puts it in the *Critique of Pure Reason*, Leibniz "took the appearances for things in themselves" (1929 [1781/7], A264/B320). (The customary practice of

citing the pagination of the 1781 (A) edition and/or the 1787 (B) edition is followed.)[10] Leibniz's reduction of physical phenomena to monadic things in themselves arose from his reduction of relations, and relational causal powers, to the intrinsic properties of monads. The forces of Leibniz's dynamic physics were in the end nothing over and above the intrinsic properties of monads. That is what Kant rejects. According to Kant, endowing substances with real connections, or real powers to connect, involves more than making individual substances with their intrinsic properties – it involves an additional act of creation.

What, if anything, follows from this contingency, for Kant? In *Kantian Humility* (Langton 1998) I argued that intrinsic properties are, for Kant, causally inert: that is why Kant says "it is never through its intrinsic properties that a substance has the power to determine others different from itself" (415). I also argued that this assumption persists through Kant's intellectual career. It reappears in the *Critique of Pure Reason*, in the context of a discussion of Leibniz:

> Monads are supposed to serve as the raw material for the whole universe, despite having no active force except for that consisting in representations (which, strictly speaking, are active only within the monads). That is why Leibniz's principle of the possible reciprocal community of substances had to be a pre-established harmony, and not a physical influence. For when everything is merely intrinsic . . . the state of one substance cannot stand in any active connection whatsoever with the state of another. (A274/B330)

The contingency Kant first identified in the *New Exposition* remains a continuing theme, and provides the key – or so I argued – to Kant's conclusion that we are ignorant of things in themselves.[11]

The modal mistake Kant saw in Leibniz has an epistemological side. The supposed necessary connection between phenomena and things in themselves offers an epistemological route from phenomena to things in themselves. If we can "take" appearances to be nothing over and above the things in themselves, then perceptual acquaintance with appearances yields acquaintance (perhaps "confused" acquaintance) with things in themselves, in the way that acquaintance with a crowd is acquaintance with the people that make up the crowd, or acquaintance with the roar of the ocean is acquaintance with the individual waves that make up that roar. Acquaintance with matter, constituted by the physical forces of Leibniz's dynamics, is acquaintance with the living forces intrinsic to the monads.

If that necessary connection goes, the bridge falls: knowledge of phenomena does not yield knowledge of things in themselves. If relational causal powers are something different to the substances as they are "in themselves," then knowledge of relational causal powers no longer yields knowledge of things as they are in themselves. That is what Kant means when he says, in the Transcendental Aesthetic:

> Everything in our knowledge which belongs to intuition . . . contains nothing but mere relations, of locations in an intuition (extension), of change of location (motion) and of laws according to which this change is determined (moving forces). What presents itself in this or that location, or, beyond this change of location, what activities occur within the things themselves, is not given through these relations. *Now through mere relations one cannot be acquainted with a thing as it is in itself.* We may therefore conclude that since external sense gives us nothing but representations of mere relations, this sense can contain in its representation only the relation of an object upon the subject, and not the intrinsic properties that belong to the object as it is in itself. (B67, emphasis added)

More needs to be said, but perhaps that is enough to suggest how the contingency in Kant describes in his early work might, in the end, bear on ignorance of things in themselves.

31.1.2 Lewis

David Lewis argued in "Ramseyan Humility" that, whether or not Humility is right as an account of Kant, it is simply right. The idea that "we find out nothing about [things] as they are in themselves," is "true . . . or at least something very like it is." Kantian ignorance of things in themselves is, he says,

> ignorance of the intrinsic properties of substances. The substances that bear these intrinsic properties are the very same unhidden substances that do indeed affect us perceptually. But they affect us, and they affect other things that in turn affect us, in virtue of their causal powers, which are among their relational properties. Thereby we find out about these substances as bearers of causal powers, but we find out nothing about them as they are in themselves.[12]

Things affect other things in virtue of their causal powers, but the things "in themselves" remain hidden.

Lewis notes that the Kantian argument relies on an assumption about the *inertia* of intrinsic properties, which many would find implausible. Suppose we deny it, and say more plausibly that it is precisely in virtue of their *intrinsic* properties that things affect other things. Would the grounds for Humility disappear? No they would not – because the connection between intrinsic and relational properties remains contingent. This, I argued – and Lewis agrees – still leaves the intrinsic properties as a something-we-know-not-what. As Lewis continues, "to be the ground of a disposition is to occupy a role, but it is one thing to know that a role is occupied, another thing to know what occupies it" (2009, 204).

He spells this out in terms of a contingent relationship between a role and what realizes that role.

> Being the ground of a certain disposition is only one case among many of role occupancy. There are a variety of occupied roles, among them nomological roles and others as well. Quite generally, to the extent that we know of the properties of things only as role-occupants, we have not yet identified those properties. No amount of knowledge about what roles are occupied will tell us which properties occupy which roles.[13]

Different properties could, says Lewis, occupy the very same role.

His argument proceeds from Ramseyan premises about how theoretical terms get their meaning. Imagine science were to give us a grand, "final theory" of the world, says Lewis; a theory T which would yield the fundamental, intrinsic properties that play an active role in the workings of nature. The language of this theory contains terms implicitly defined by the theory. The theory has a unique actual realization, and the terms of the theory name fundamental properties. The theory leaves out what Lewis calls "idlers," fundamental properties (if any) that are actually instantiated but play no active role. It also leaves out "aliens," fundamental properties that are not actually instantiated.

Besides our grand, final theory, there is the rest of our language, our "old" language, call it "O," which happens to be rich enough to express our observations. We get the Ramsey sentence of T by replacing the names in T with existentially quantified variables. The Ramsey sentence says that T has at least one actual realization. It implies the O-language sentences that are theorems of T. Any predictive success for the theory T is also a predictive success for the Ramsey sentence. If the theory T has more than one possible realization, observation won't help us know which realization is actual, because no possible realization gives us evidence that goes beyond the Ramsey sentence. If our theory

T has more than one possible realization, then there are some fundamental properties that remain hidden from us.

Does T have more than one possible realization? Yes – because of the contingency of the connection between relational role properties and intrinsic realizer properties. Lewis argues for it in two ways. He offers a "permutation" argument: two fundamental properties named by T could be permuted, leaving everything else just the same, in which case T would have a different realization. He offers a "replacement" argument: the fundamental properties named by T could be replaced by "idler" or "alien" properties, in which case again T would again have a different realization. So our theory has different possible realizations, and we can't tell which of these is actual.

This sketch doesn't do justice to Lewis's initial argument, still less to its later developments, but the basic idea can be expressed in terms of a familiar picture from functionalist philosophy of mind.[14] Just as our concept of "pain" gets its meaning from its place in our theory about how pain is related to other mental states and behavior, so in physics our concept of "positive charge" gets its meaning from its place in our theory: things having positive charge attract things having negative charge, and so on. "Pain" contingently refers to a neural state, and the role-property of pain is on this way of thinking multiply realizable, by neurons, circuit boards, ectoplasm. Likewise, then, "electron" contingently refers to whatever actually realizes the role of electrons, and the role property of being an electron is multiply realizable by . . . what? All we can say is: different possible realizer properties. When it comes to our theory of mind, we can dig into other theories to give us a sense of what the different possible realizers of role properties might be; but when it comes to our theory of the physical world, we can dig no deeper. We know what realizes the pain role, for us, namely certain neural states. But what realizes the positive charge role? We can only shrug.

31.1.3 Kant and Lewis Compared

There are differences between the Kantian and Lewisian pictures, and not only in their supplementary premises. We are, as advertised, skating over the differences, but let us briefly consider their responses to the question why intrinsic properties are needed in the first place. On the Kantian proposal, intrinsic properties are needed for the existence of a substance that is an ultimate subject. As Kant puts it in *Physical Monadology*, "besides external presence, i.e. relational determinations of substance, there are other, internal, determinations, without which the relational determinations would not be, because there would be no subject in which they inhered."[15] This notion of substance matches, I have argued, the "pure concept" of substance described in the *Critique* (e.g. B307, A284/B340, A274/B330).[16] Intrinsic properties are needed because an independent thing needs independent properties.

Why are intrinsic properties needed for Lewis? They are needed as realizers for certain role properties, as the categorical bases for certain dispositional properties, and, for Lewis, the fundamental intrinsic, or categorical, properties remain causally potent, although unknown. The requirement that roles have realizers, or that dispositional properties have categorical bases, is not the same as a requirement that substances have some intrinsic properties or other.[17]

Notwithstanding these differences, Kant and Lewis have something in common. They agree that we are ignorant of things in themselves. They agree that there exist certain intrinsic properties, and we don't know what they are. They agree that this ignorance is due at least in part to a contingent connection between two classes of properties, deserving of the labels "intrinsic" and "relational." Putting the contingency in metaphorical form: for Kant, God could make the intrinsic facts one way, and the relational facts a different way. God could make facts about substances and their intrinsic

properties one way, and facts about their causal powers a different way. For Lewis, God could make the relational facts one way, and the intrinsic facts a different way: he could make the facts about role properties one way, and facts about realizer properties a different way. Both ways, the contingency in question – perhaps with help from supplementary premises – renders the intrinsic properties inaccessible. For both, contingency leads to Humility.

To what extent are supplementary premises required? A premise, say, about the *receptivity* of knowledge, for Kant: must we add, for Kant, the idea that in order to have knowledge of something we must be affected by it? (That is what I originally argued.) And a premise, say, about the Ramsified meaning of theoretical terms, for Lewis? The question about how much these arguments depend on further premises is interesting and important, but I have enough on my plate as it is.[18]

31.1.4 Escape from Humility?

Supposing that we *are* ignorant of things in themselves, there is a question of how gloomy to be. Kant thinks we are missing out on something – he speaks of our yearning for something more, of doomed aspirations, of "our inextinguishable desire to find firm footing somewhere beyond the bounds of experience" (A796/B824). Lewis agrees we are missing out, and is less gloomy: whoever promised me I could know everything? But, if you are gloomy, you might want a remedy.

One remedy might be to deny the contingency. We needn't deny it in the way that Leibniz did, by affirming a necessary connection, taking "the appearances for things in themselves" – taking the world of relational, causal powers constituting matter to be nothing over and above the world of monads. We could deny the contingency a different way.

Suppose we deny a gap between relational roles and intrinsic realizers. We could say that the role a property plays is not, after all, something contingently worn, not a cloak that can be thrown off (so to speak) when travelling from world to possible world. Sydney Shoemaker, for example, proposes that

> what makes a property the property it is, what determines its identity, is its potential for contributing to the powers of things that have it . . . if under all possible circumstances properties X and Y make the same contribution to the powers of the things that have them, X and Y are the same property. (1980, 256)

On this account, nomological role is essential to a property: playing the positive charge role, for example, is essential to whatever property it is that realizes positive charge. What plays the role of positive charge could not fail to play that role. This account of properties brings a kind of necessity into nature that is absent on Lewis's picture, and absent on a traditional account of dispositions. It may have implications for the modality of laws of nature: perhaps laws are no longer contingent, but metaphysically necessary (see also Ellis and Lierse 1994; Hawthorne 2001; Bird 2007). It may have implications for our assumptions about intrinsicality: if these fundamental properties are essentially tied to their nomic role, perhaps they are no longer intrinsic. These conundrums I'll flag, and leave aside. But the upshot of this denial of contingency may indeed be an escape route from Humility – or at any rate, an escape from *Ramseyan* Humility, as Lewis himself concedes.

How, if at all, does this bear on Kantian Humility? If causal structuralism puts necessity into nature by building a necessary connection between powers and categorical properties, Kant in the end does something else. In place of a metaphysically necessary connection between classes of properties, categorical and dispositional, he argues for a "transcendentally" necessary connection between experience and its objects – but this "transcendental" necessity, in contrast to the necessity sought by causal structuralism, provides no escape route from Humility, as we shall see.

31.2 Necessity and Coexistence

31.2.1 Kant

The goal of Kant's argument in his 1755 *New Exposition* was not simply to establish the contingency we have just been considering, but to establish what he called a "Principle of CoExistence."

> Finite substances are, through their solitary existence, unrelated, and are evidently not connected by interaction, except in so far as they are maintained by the principle of their common existence (namely the divine intellect) in a pattern of mutual relations. (1986, 413)

According to the Principle of Coexistence, substances do not coexist in the same world unless they bear "mutual relations" – causal relations, such as the physical forces of Newtonian attraction – towards each other. Having the power to enter into such relations is therefore necessary, but presumably not sufficient, for coexistence. (Here we again run up against the unresolved question of whether Kant has relations or relational powers in mind – or whether indeed he failed to distinguish these sufficiently well.) This structure of mutual relations is not fixed by the substances and their intrinsic properties but is "added" to the substances in a special divine creative act. However we interpret Kant's argument, he evidently rules out the idea that coexistence, the relation that enables things to be "worldmates" (Lewis's phrase), might be simply primitive, and he identifies these "mutual relations" as a certain set of physical relations. You and I inhabit the same world, it seems, in part because the forces of gravity hold us together! – a peculiar, but somehow reassuring, thought.

Kant returns to this topic in the *Critique of Pure Reason*, where his interest is no longer in the metaphysical conditions for coexistence, but rather in the conditions for our *experience* of coexistence. Having chastened the ambitions of dogmatic metaphysics, Kant is now pursuing a different methodology, that of the "transcendental" argument. Broadly speaking, a transcendental argument begins with some ordinary and obvious feature of our thought, or experience, and then moves on to identify some surprising, far from obvious, necessary conditions of that thought, or experience. It is an understatement to say that Kant put transcendental arguments on the philosophical map; and what we find in the Third Analogy is a transcendental argument for a Principle of Community, the later incarnation of Kant's early Principle of Coexistence.

> All substances, in so far as they can be perceived to co-exist in space, are in thorough-going reciprocity . . . Now suppose that a plurality of substances as appearances were each completely isolated – so that none would act on the other, and would receive from the other no reciprocal influence. I say that the co-existence of these substances would be no object of possible perception, and that the existence of one of them could not lead through the way of empirical synthesis to the existence of the other . . . Each substance must therefore contain in itself the causality of certain properties in the others, and at the same time contain in itself the effects of the causality of others, that is, they must stand in dynamical community if their co-existence in a possible experience is to be known. Now, with respect to objects of experience, something is *necessary* if experience of the objects themselves would be impossible without it. Therefore it is necessary that substances in the [field of] appearance, in so far as they are co-existent, should stand in a thorough-going community of mutual interaction. (A213/B260)

Kant again wants to affirm the necessity of "mutual relation," with respect to coexistence, but his question now is not about the conditions for a world of coexisting objects, but rather what the conditions for *experience* of a world of coexisting objects. Forget God. Forget divine acts of creation. What must we *take* the world to be like, if we are to *experience* ourselves as part of a world, coexisting with other parts?

The passage is not an easy one, but perhaps we can see Kant's answer as proceeding by *reductio*. If there were a plurality of substances "as *appearances*" each of which was completely isolated from the others, then they would not act upon each other, and (contrary to the hypothesis) would not be objects of possible perception, and would not be substances "as appearances." Hence substances as appearances must not be isolated, but stand in dynamical community. In the background is the early idea that independence of substance implies contingency: that isolated, or isolable, substances do not have what it takes to get the causal power that would allow them to interact, and thereby to coexist. The mere existence of substances, each capable of existing by itself, is not sufficient for interaction among them. But Kant now gives up on that problem. He abandons the question of how a substance might be endowed by a creator with relations in order to form a world. He affirms the contingency of his early *New Exposition*, and makes it irrelevant. The point is that such isolated (or isolable) substances would not be objects of possible perception: if we are to have experience of things coexisting with us, and each other, at the same time, we must already be (or *think* of ourselves as being) in a world of causally interacting substances.

A comparable argument occurs in a different part of the *Critique*, the Postulates of Empirical Thought. Here again, the demand for experienced community – for experience of ourselves as members of a world in which we coexist with other substances – is seen to generate a demand for physical causal relations of "action and reaction," and Kant says this is made possible only by representing these substances in space.

> How are we to think it possible that, when several substances exist, from the existence of one, some reciprocal action on the existence of the others can follow? We think that because something is in the one there must also be something in the others which is not to be understood solely from their own existence alone. For that is what community requires. But among things that are each totally isolated (each isolated in its own subsistence) community is completely inconceivable. . . . However we can make quite comprehensible the possibility of a community of substances – substances as *appearances* – when we represent them to ourselves in space, in external intuition. For this already contains in itself *a priori* formal external relations as conditions of the possibility of real external relations (action and reaction) and thus community. (B293–4)

He argues that space provides the formal relations that enable us to represent the real dynamical relations of action and reaction, and thus to represent substances as coexisting appearances.

I'm not defending Kant's argument; just emphasizing that for Kant, coexistence is no primitive, but requires grounding in physical, causal relations. This remains a continuing theme, whether he is arguing for the metaphysically necessary conditions for coexistence in his early writings, or the transcendentally necessary conditions for our experience of coexistence, in the *Critique of Pure Reason*.

31.2.2 Lewis

At the beginning of *On the Plurality of Worlds*, Lewis asks a question which could have come from Kant's 1755 *New Exposition*, about the conditions under which individuals could coexist and form a world.

> What, then, is the difference between a sum of possible individuals that is a possible world, and one that is not? What makes two things worldmates? How are the worlds demarcated from one another? Why don't all the possibilia comprise one big world? Or at the other extreme, why isn't each possible neutrino a little world of its own? (1986, 70)

Lewis offers an answer rather like Kant's: a system of real, physical relations provides the ground of coexistence.

> [N]othing is so far away from us in space, or so far in the past or the future, as not to be part of the same world as ourselves . . . So we have a sufficient condition: if two things are spatiotemporally related, then they are worldmates. The converse is much more problematic. Yet that is more or less the doctrine that I propose. Putting the two halves together: things are worldmates iff they are spatiotemporally related . . . There are no spatiotemporal relations across the boundary between one world and another; but no matter how much we draw a boundary within a world, there will be spatiotemporal relations across it. (1986, 70–1)

For Lewis, physical, spatiotemporal relations are the glue that binds individuals into one world. In Kant's *New Exposition*, that work is done by "mutual relations" that are causal (not, in the first instance, spatial), including the actions of such forces as gravity. Lewis too sees a need for relations of causal connectedness as well as spatial connectedness – but thinks he needn't mention the former, since he gets them for free.

> There is a second way in which the worlds are isolated: there is no causation from one to the another. If need be, I would put this causal isolation alongside spatio-temporal isolation as a principle of demarcation for worlds. But there is no need. Under a counterfactual analysis of causation, the causal isolation of worlds follows automatically. (1986, 78)

Kant, in the *New Exposition*, has it the other way around, causal relations as prior to spatial ones. Kant thinks at this early stage that he gets *space* for free – that spatial relations supervene on the dynamical. He would have detested Lewis's reductionism about causation, causal power, and law.

All the same, we have arrived at a curious destination: Lewis's version of Kant's early principle, the Principle of Coexistence. That's if we construe Lewis as offering an argument in what Kant came to call "dogmatic metaphysics," for the conclusion that it is metaphysically necessary that worldmates bear real physical relations to each other. Lewis's version is not quite Kant's; for Lewis there is no special role for reciprocal, symmetrical causal relations. But we still have a conclusion that the early Kant would applaud: I am spatiotemporally and causally connected with all the other things in my world; otherwise we would not even coexist.

What of Kant's later Principle of Community, in the *Critique of Pure Reason*? On the face of things, Lewis would have no part of it. Lewis is a Goliath of metaphysics – not of transcendental argument! But let's step back a moment. We said a transcendental argument begins with some ordinary and obvious feature of our thought or experience, and moves on to identify some far from obvious conditions of that thought or experience.

Well now, doesn't Lewis move from some uncontroversial features of our thought, to some far from obvious conditions of that thought? Some philosophers have suggested that we take seriously the idea of a *transcendental* argument for modal realism, and that Lewis might be construed in these very terms.[19] We begin with a routine feature of our thinking. We think that "things might have been different in ever so many ways," as Lewis puts it in the opening paragraphs of *On the Plurality of Worlds*.

> This book of mine might have been finished on schedule. Or . . . I might be defending not only a plurality of possible worlds, but also a plurality of impossible worlds . . . Or there might never have been any people. Or the physical constants might have had somewhat different values, incompatible with the emergence of life. (1986, 1)

We *think modally*: we think some things could have been different, others couldn't have been. It then turns out, surprisingly, that this modal thinking is *itself* possible only if there is (or we think there is) a plurality of existing possible worlds, whose actuality is a merely indexical matter. Our world is actual, but different from the others only in being ours. Forcing Lewis's argument into Kantian idiom: a necessary condition of the possibility of modal thinking turns out to be a commitment to real possible worlds. So we can see Lewis as offering an argument about the transcendental necessity of modal realism itself.

Modal realism in turn requires that worlds be individuated: and this is what brings us to Lewis's version of the Principle of Community. When Lewis asks about the conditions under which individuals coexist and form a world, we can shift our perspective and see him as asking a transcendental question about the conditions of our modal thinking. The conditions for modal thought include a commitment to modal realism – which in turn brings a commitment to a distinctive, physical account of the relation of coexistence. Thinking modally requires us to think that real physical relations provide the grounds for coexistence: we must think that

> things are worldmates iff they are spatiotemporally related . . . There are no spatiotemporal relations across the boundary between one world and another; but no matter how much we draw a boundary within a world, there will be spatiotemporal relations across it. (Lewis 1986, 70–1)

Thinking modally likewise requires us to think (though we get this for free) that things are worldmates just in case they are *causally* related with each other as well.

Well now, another curious destination. We began with an uncontroversial premise about our modal thinking, and reasoned to some highly non-obvious conditions of that thinking, namely modal realism itself; which in turn requires us to think we are in causal and spatiotemporal community with all the individuals with which we coexist. If that's not a transcendental argument, I'm not sure what is. We earlier found Lewis's version of Kant's early Principle of Coexistence, and we've found an analogue to his Principle of Community, in the Third Analogy. Again, not quite the Principle of Community; for example, Lewis has no special role for reciprocal, symmetrical causal relations. But we still have a conclusion that the later Kant might applaud: a necessary condition of the possibility of thought (*modal* thought, in this case) includes a commitment to spatiotemporal and causal connectedness among coexisting things. My worldmates and I must occupy the same spatial and causal nexus, to be worldmates at all.

31.3 Conclusion

At the end of section 31.1, we considered how one could avoid Humility by moving to causal structuralism. Properties are identified with their causal profile, and necessarily interact as they do: so that necessarily, if something has positive charge, it attracts something that has negative charge. We've now considered how Kant moves to a different kind of necessity, a transcendental necessity, about many things, including the conditions for experience of coexistence in one world. And we've explored the antecedently unlikely prospect of a transcendental argument on Lewis's behalf, for a Principle of Community comparable to Kant's.

Connecting some of the dots now: one might wonder whether Kant's transcendental necessity does a job comparable to the necessity in causal structuralism. If contingency was the problem, do these affirmations of necessity *both* offer a solution, a way to escape or soften the Humility that is otherwise our fate? No. Kant has transcendental necessity – *and* Humility. The problem is that for

Kant, unlike the causal structuralist, the original contingency still holds. He said in the *New Exposition* that "it is never through its intrinsic properties that a substance has the power to determine others different from itself," and he continues to believe it, reaffirming the same absence of metaphysical necessity in the *Critique* – "when everything is merely intrinsic … the state of one substance cannot stand in any active connection whatsoever with the state of another" (A274/B330).

Whatever the mutual causal relations might be that enmesh us in the world, it is not through the things as they are in themselves – not through their intrinsic properties – that those relations hold. So for Kant, the transcendental necessity of the Third Analogy is compatible with the metaphysical contingency of the *New Exposition*. The transcendental necessity of our experiencing a causally interactive world is compatible with a metaphysical contingency – things in themselves fail to necessitate causal power.

So some ways of denying contingency may save knowledge, and others may not. Causal structuralism's denial of contingency may forestall Lewis's Ramseyan ignorance of things in themselves. But transcendental philosophy's denial of contingency will not forestall Kantian ignorance of things in themselves. For Kant, the transcendental necessity of the Third Analogy does nothing to overcome the contingency that puts things in themselves beyond our grasp. The *Critique's* efforts to save knowledge do not get us knowledge of everything; we are left, still, with "no insight into the intrinsic nature of things" (A277/B333).

Notes

1 Langton 1998. This "contingency" was discussed there under the heading of the "irreducibility" of relational properties to intrinsic properties.

2 Desmond Hogan has also argued, differently, for a connection between contingency and ignorance of things in themselves, in Kant. He thinks such ignorance arises from the non-entailment of certain features by "determining grounds" in things in themselves, because of the *freedom* of the thing in itself, see Hogan (2009).

3 Lewis 2009, 203. He says "that of Kant – *according to Langton*" (my emphasis), observing that my interpretation is controversial.

4 Ralf Bader has commented (in correspondence) that even though contingency plays a role in Lewis's permutation and replacement arguments, discussed here, these arguments seem to drop out of the picture once one brings in the idea, emerging later in Lewis's paper, that our ignorance is "ineffable." I do not do justice to this excellent point here.

5 Kenneth Walden prompted me to think about the possibility of transcendental arguments in Lewis: Walden regards Lewis's indispensability argument for modal realism as a transcendental argument, or at any rate as close as a post-Quinean will ever get to one. See Walden 2010. Others who at least take seriously a transcendental argument for modal realism include Shalkowski 1994, Hawthorne 1996, and esp. Bricker 2006. I am grateful to Kenneth Walden and Antony Fisher for helpful conversation on this topic.

6 See, for example, *Physical Monadology* (1986b [1756]) and, for more detailed references, *Kantian Humility* (Langton 1998, esp. chapters 3 and 5). For expression in later work, see, for example, the *Critique of Pure Reason* (1929 [1781]): Transcendental Aesthetic and the Amphiboly of the Concepts of Reflection.

7 Lloyd Humberstone (1996, 205–67) advises using "intrinsic/extrinsic" to label metaphysically robust properties, "non-relational/relational" to label what he calls "property-concepts" – good advice, which I nonetheless don't live up to here.

8 For an amendment on these lines see Langton (1998, chapter 5), which is comparable to an option considered in Langton and Lewis (1998). In the latter paper we aim to improve the isolation test. We propose that a property is a *basic intrinsic* property iff it is *independent* of loneliness and accompaniment: its presence is compatible with isolation, and with accompaniment; so too is its absence. All other intrinsic properties supervene on the basic intrinsic ones. A further interpretive issue I set aside concerns the "mutuality" of

these relations or powers, understood as somehow symmetrical or reciprocal: see Eric Watkins, *Kant and the Metaphysics of Causality* (2005) for excellent discussion.

9 See again Watkins 2005, for sustained development of the argument that Kant has considerably more than event causation in mind, both here and in the Critical period.

10 See Langton 1998, chapter 4.

11 See also Kant 1929 [1781] (A213/B260; B293–4), and for more references, Langton 1998, 133–8.

12 Lewis 2009, 203.

13 Lewis 2009, § 1.

14 This analogy is a theme in Robichaud 2010.

15 Kant 1986a, Prop. VII.

16 See Langton 1998, chapter 3.

17 Even setting aside the distinction between the intrinsic and the categorical, which demands more attention than we're giving it here: for more discussion of this see Langton and Robichaud 2010.

18 Note in this connection Ralf Bader's comment that contingency plays no role in Lewis's later argument about "ineffability" in "Ramseyan Humility" (pers. comm.).

19 See again Walden 2010; Shalkowski 1994; Hawthorne 1996; and esp. Bricker 2006.

References

Bird, A. (2007). *Nature's metaphysics: laws and properties*. Oxford: Oxford University Press.

Bricker, P. (2006). Absolute actuality and the plurality of worlds. *Philosophical Perspectives*, 20, 41–76.

Ellis, B. & Lierse, C. (1994). Dispositional essentialism. *Australasian Journal of Philosophy*, 72, 27–45.

Hawthorne, J. (1996). The epistemology of possible worlds: a guided tour. *Philosophical Studies*, 84, 183–202.

Hawthorne, J. (2001). Causal structuralism. *Philosophical Perspectives*, 15, 361–78.

Hogan, D. (2009). How to know unknowable things in themselves. *Noûs*, 43, 49–63.

Humberstone, L. (1996). Intrinsic/Extrinsic. *Synthese*, 108, 205–67.

Kant, I. (1986a [1755]). *A new exposition of the first principles of metaphysical knowledge*. In L.W. Beck (Ed.), *Kant's Latin writings: translations, commentaries and notes*, pp. 57–109. New York: P. Lang. Translation adapted from Academy edition of Kant's Gesammelte Schriften. Berlin and Leipzig: de Gruyter, 1922.

Kant, I. (1986b [1756]). *Physical monadology*. In L.W. Beck (Ed.), *Kant's Latin writings: translations, commentaries and notes*, pp. 473–88. New York: P. Lang. Translation adapted from Academy edition of Kant's Gesammelte Schriften. Berlin and Leipzig: de Gruyter, 1922.

Kant, I. (1929 [1781/7]). *Critique of pure reason*. Trans. N. Kemp Smith. London: Macmillan. The customary practice of citing the pagination of the 1781 (A) edition and/or the 1787 (B) edition is followed. Translation adapted from Academy edition of Kant's *Gesammelte Schriften*. Berlin and Leipzig: de Gruyter, 1922.

Langton, R. (1998). *Kantian humility: our ignorance of things in themselves*. Oxford: Oxford University Press.

Langton, R. & Lewis, D. (1998). Defining 'intrinsic'. *Philosophy and Phenomenological Research*, 58, 333–45.

Langton, R. & Robichaud, C. (2010). Ghosts in the world machine? Humility and its alternatives. In A. Hazlett (Ed.), *New Waves in Metaphysics*, pp. 156–78. New York: Palgrave Macmillan.

Lewis, D. (1986). *On the plurality of worlds*. Oxford: Blackwell.

Lewis, D. (2009). Ramseyan humility. In Nola & Braddon-Mitchell (2009), pp. 203–22.

Nola, R. & Braddon-Mitchell, D. (2009). *Conceptual analysis and philosophical naturalism*. Cambridge, MA: MIT Press.

Robichaud, C. (2010) What lies beneath? A defense of categorical humility. PhD dissertation, MIT.

Shalkowski, S. (1994). Ontological ground of alethic modality. *Philosophical Review*, 103, 669–88.

Shoemaker, S. (1980). Causality and properties. In P. van Inwagen (Ed.), *Time and cause*, pp. 109–35. Dordrecht: Reidel.

Walden, K. (2010). Lewis, Kant and the Self. In M. Guillot (Ed.), *Working papers No. 5: self-locating beliefs*, Paris, March 25–26. Institut Jean Nicod/MIT Department of Linguistics and Philosophy, http://cpr.nicod.free.fr/Texte/W5.pdf, accessed June 19 2014.

Watkins, E. (2005). *Kant and the metaphysics of causality*. Cambridge: Cambridge University Press.

32

Analytic Functionalism

WOLFGANG SCHWARZ

32.1 Overview

Lewis held that psychological states are individuated by their causal-functional profile. Pain, for example, may be characterized as whatever state is typically caused by burns and injuries, causes such-and-such signs of distress, a desire for the state to go away, and so on. If it turns out that some biological state, say C-fiber firing, uniquely plays this role, then it has turned out that C-fiber firing is pain. According to Lewis, the roles that characterize mental states can be extracted from *folk psychology*: our tacit but shared beliefs about how mental states interact with one another, what kind of behavior they tend to cause, and how they change under the impact of perceptual stimulation. Folk psychology implicitly defines our mental vocabulary: "pain," "hunger," etc. *mean* "whatever state plays this and that role." Thus psychological truths are analytically entailed by non-psychological truths. If you know what "pain" means, and you know that C-fiber firing plays the relevant role, then you can infer with certainty that people whose C-fibers fire are in pain.

Lewis's position, often called *analytic functionalism*, was inspired by Ryle's (1949) analytic behaviorism, which took psychological predicates to express complex sets of behavioral dispositions. On this view, to say that someone is hungry is to say that they would eat if offered food, that they would more likely go to a restaurant than to a bar, etc. Lewis instead identifies hunger with the inner state that provides the causal basis for these dispositions. (Lewis wasn't the first to make this proposal, see e.g. Reichenbach 1938.) This vindicates the use of psychological predicates in causal explanations: you went to the restaurant *because* you were hungry; your hunger is part of what caused you to choose the restaurant over the bar. It can also accommodate the fact that mental states often have distinctive behavioral effects only in combination with other mental states. A desire for happiness can manifest itself in all kinds of behavior, depending on what else the subject believes and desires. In addition, analytic functionalism can allow that paralyzed people (say) may have pain or hunger even though they lack the relevant behavioral dispositions. That's because folk psychology is full of *ceteris paribus* clauses: pain and hunger are defined as states that *typically* ground such-and-such dispositions.

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

In this chapter, I will review some of the main tenets of Lewis's philosophy of mind. I will begin with some comments on the methodology Lewis employed in his analysis of psychological terms, which has become standard in functionalist accounts across philosophy. Then I discuss the difference between what are often called "realizer functionalism" and "role functionalism," and argue that Lewis made the wrong choice. Section 32.4 presents Lewis's often misunderstood account of intentionality. In section 32.5, I end with a few pessimistic remarks on the prospect of analyzing phenomenal truths in terms of functional roles.

32.2 The Canberra Plan

Lewis subscribed to the familiar Fregean view that the meaning (in some sense of "meaning") of referential expressions is given by conditions that specify their reference under hypothetical circumstances. The conditions are common knowledge among fully competent members of the relevant linguistic community, and elicited by intuitions of the kind familiar from Kripke and Gettier. Often the reference conditions associated with a term can be made explicit by looking at a certain body of statements – a "theory" – in which the term occurs. For the term "entropy," the relevant theory would be thermodynamics or statistical mechanics. For more ordinary terms like "pain" or "water," the right starting point is our folk theory, "a generally shared body of tacit belief" about the relevant subject matter (Lewis 1997b, 333). The folk theory of water might say that water covers a large part of the Earth, quenches thirst, is typically transparent, and so on.

Imagine such a theory written as a single sentence *T*. The *matrix of T* is the same sentence with all occurrences of the relevant term (or terms) replaced by variables (different variables used for different terms). The matrix of the above water theory, for example, would begin with "*x* covers a large part of the Earth and *x* quenches thirst," and so on. This expresses a condition, or *role*. If an entity satisfies the condition, it is called a *realizer* of the role. The chemical substance H_2O is arguably a realizer of the water role, because H_2O covers a large part of the Earth, quenches thirst, and so on. If the matrix of *T* contains more than one free variable, then a realizer is a list of entities rather than a single entity.

Existentially binding the free variables in the matrix of *T* yields the *Ramsey sentence of T* (see Ramsey 1931b). The Ramsey sentence of the water theory says that there is an *x* such that *x* covers a large part of the Earth, etc. *T* is logically stronger than its Ramsey sentence, because it doesn't only say that *something* is so-and-so, but that *water* is so-and-so. Thus the water theory is false and its Ramsey sentence true if something other than water plays the water role, while water itself does not play that role. However, if "water" is implicitly defined by the water role – that is, if the reference conditions for "water" are given by the matrix of the water theory – then this possibility can be ruled out: it could not turn out that something plays the water role, but water doesn't. Hence the Ramsey sentence of *T* is *a priori* equivalent to *T*. No possible discovery could reveal that one of them is true and the other false. The non-empirical part of *T* that logically goes beyond the Ramsey sentence is captured by the *Carnap conditional of T*: the material conditional with the Ramsey sentence of *T* as antecedent and *T* itself as consequent (see Carnap 1963). The Carnap conditional isolates the analytic, definitional component of the theory.

People sometimes object that there are not enough analytic truths about things like water to make this story work: surely it isn't analytic that water covers a large part of the Earth, or that it quenches thirst. That is true, but irrelevant. The individual conjuncts of the water need not be analytic. What's supposed to be analytic is only the Carnap conditional of the entire theory. In general, the more empirical claims are added to a theory, the harder it gets to falsify its Carnap conditional. Recall that, in order to falsify a material conditional, one has to find that the antecedent is true and the

consequent false. Could it turn out that something plays the water role, but water does not? If the water role is merely to cover a large part of the Earth, it is easy enough to imagine some such discovery. On the other hand, suppose the water theory contains absolutely everything you believe about water. Could it turn out that all these things are true of *something*, while they are not true of water? How could you discover that this something isn't water if there is no feature which you think water has but this other substance lacks?

The downside of adding a lot of empirical assumptions to a theory is that it may then easily turn out that there is no actual realizer of the theoretical role. The Carnap conditional then remains true, but it becomes useless for locating the relevant phenomenon in fundamental reality. We can't identify water with the chemical kind that realizes the water role if nothing realizes that role. But we also wouldn't say that water does not exist if it turns out that some of our water beliefs are false. We know that most of our theories about most things may well be mistaken, even in quite central respects. The matrix of our theories is therefore too strong to express the reference conditions for the relevant terms. Lewis's usual response is to say that things may count as referents of a term even if they fall somewhat short of realizing the associated role (see Lewis 1984, 59; 1994, 298; 1995; 1996, 58; 2004b, 280). In Lewis (1966, 104) and (1970, 83) he makes a slightly different suggestion, which I like better: to weaken the theoretical roles. Let me illustrate a natural way how this can be done.

Let's begin with the Carnap conditional of our total water theory: if something satisfies all our water beliefs, then surely that something is water. What if nothing satisfies all our beliefs, but something satisfies everything except that it does not occur on Mars, or in cucumbers? Then that something is still water. At some point, as we drop or revise more and more of the original assumptions, it becomes unclear whether the thing that realizes the revised role still deserves the name "water." If nothing comes even close to realizing the water role, then there is no water. This is how scientists once discovered that phlogiston or the planet Vulcan do not exist: they found that nothing comes even close to playing the role associated with those terms. Now we have a list of Carnap-style conditionals, with increasingly weakened or modified versions of the original theory T:

$$\exists x T(x) \supset T(\text{water}),$$
$$\big(\neg\exists x T(x) \wedge \exists x T'(x)\big) \supset T'(\text{water}),$$
$$\big(\neg\exists x T(x) \wedge \neg\exists x T'(x) \wedge \exists x T''(x)\big) \supset T''(\text{water})$$

The conjunction of these conditionals is logically equivalent to the single Carnap conditional

$$\exists x T^*(x) \supset T^*(\text{water}),$$

where $T^*(x)$ is defined as

$$T(x) \vee \big(\neg\exists y T(y) \wedge T'(x)\big) \vee \big(\neg\exists y T(y) \wedge \neg\exists y T'(y) \wedge T''(x)\big) \vee \ldots.$$

$T^*(x)$ is the weakened theoretical role that does a better job at capturing the reference conditions for the relevant term than $T(x)$. Unlike $T(x)$, it also takes into account hypothetical situations in which our theory is false. If nothing realizes the weakened water role, then there is no water.

What if a role is realized by *several* things? In early works like (1970, 83) and (1972, 252), Lewis declared that the relevant term should then be treated as empty. A better response is to say that, in such a case, the term is semantically indeterminate between the different candidates, as Lewis says in (1997b, 347; 2004b, 280; and 2009, 220, fn. 9).

Let us return to psychological terms. Here the meaning-giving theory is folk psychology. Lewis sometimes suggests that the folk psychological role of mental states is a purely *causal* role. But theoretical roles don't need to be causal. In fact, Lewis himself mentions various folk psychological truths that don't concern causal roles – that toothache is a kind of pain (1972, 258), that people who have a conscious experience typically know the essential nature of their experience (1995), and that letter boxes are red (1997b).

The claim about letter boxes, Lewis suggests, might be part of a psychophysical theory that defines *red experience*. It is interesting not only because it isn't causal but also because it is clearly not shared by all competent speakers of English. According to Lewis, the relevant Carnap conditional is analytic only in a certain subgroup of the English-speaking community. Similarly, one might say – although Lewis does not say so – that various scientific findings have entered into the theoretical role of psychological terms as used by psychologists and neuroscientists. Allowing empirical scientific facts to constrain a state's theoretical role leads to an account often called *psychofunctionalism*, which is supposed to be an alternative to Lewis's *analytic* functionalism. Lewis does insist that "scientific findings that go beyond common sense must be kept out, on pain of changing the subject" (1974, 112; see also 1994, 311f.). Lewis's "subject" presumably was the reference of terms like "pain" as used outside of scientific circles. Even then, he arguably underestimates the deferential element of ordinary usage. Especially for somewhat technical terms like "trauma" or "depression," it may well be part of the folk understanding that these terms denote whatever experts say they denote. The gap between psychofunctionalism and analytic functionalism is therefore less wide than is often assumed.

Lewis never gave a concrete analysis of terms like "pain." In other cases, where he does offer an analysis – of causation, chance, conventions or moral values – he often rejects the idea that the analysis ought to precisely match our ordinary usage, or that the relevant theory must be common belief among all competent speakers. Part of the reason is that ordinary language is full of ambiguity, indeterminacy, and context-dependence, which can stand in the way of a systematic philosophical analysis. Lewis's analyses are therefore better understood as Carnapian explications: the goal is not to precisely trace our ordinary understanding of the relevant words, but to isolate a theoretically interesting core in the vicinity of our more or less unstable and indeterminate ordinary usage. As he says about moral value: "The best I can hope for is that my . . . theory lands somewhere near the middle of the range of variation and indecision – and also gives something that I, and many more besides, could be content to adopt as our official definition . . . in the unlikely event that we needed an official definition" (1989, 86f.).

What is crucial for Lewis's brand of functionalism is that the relevant terms – whether in their ordinary or in some regimented sense – are really defined by their theoretical role. The statement that *X* plays the pain role must analytically entail that *X* is pain. This ensures that the corresponding psychological truths are analytically entailed by, and thus reducible to, non-psychological truths.

Why care about analytic entailment? Many formulations of physicalism do not require analytic or a priori entailment of psychological truths by physical truths. On some accounts, it is enough if the physical and the psychological are connected by brute "metaphysical laws" or by some kind of "grounding" relation. Another popular formulation, which Lewis himself often uses, invokes supervenience or necessitation: every possible world that exactly matches the actual world in all physical respects without containing anything else also matching it in every other respect. This would allow the entailment of the psychological by the physical to be necessary *a posteriori*, like the entailment of truths about Hesperus by truths about Phosphorus. Lewis never took this possibility seriously. Following Frank Jackson, he argued that when it comes to the entailment of all truths by the fundamental truths, the difference between metaphysical and semantic or epistemic necessitation disappears. His argument is based on the "two-dimensionalist" premise that every *a posteriori* necessity is *a priori* entailed by some ordinary contingent truth about the actual world. For example, the

necessary truth that Hesperus = Phosphorus is *a priori* entailed by the contingent *a posteriori* truth that one and the same planet plays both the Hesperus role and the Phosphorus role. Now let P be the complete truth about the distribution of fundamental properties and relations in our world, and suppose for *reductio* that A is some truth for which the conditional $P \supset A$ is necessary *a posteriori*. By the above premise, there is a further fact Q such that $Q \supset (P \supset A)$ is *a priori*. But Q itself is made true by P. Hence $P \wedge Q$ is equivalent to P, and so $Q \supset (P \supset A)$ is equivalent to $P \supset A$. Since the former is *a priori*, so is the latter (see Lewis 1994, 296f.; 2002; Jackson 1998a, 93). The problem with this argument is that if truth-making is only a matter of necessitation, then the fact that Q is made true by P does not entail that $P \wedge Q$ is *a priori* equivalent to P: the link from P to Q may itself be necessary *a posteriori*. The argument from two-dimensionalism doesn't work. It could be repaired by adding a further premise to the effect that the fundamental truths are "semantically stable" so that primary and secondary intension coincide, but it is not clear to me whether Lewis would have endorsed this premise. At any rate, Lewis's ambition was to show how psychological truths are *a priori* or analytically entailed by physical truths – merely "metaphysical" connections are not enough.

This form of reductionism presupposes an analytic–synthetic distinction, but only a comparatively mild form. Remember that Lewis accepts that ordinary usage is often too shifty and indeterminate to allow for precise tracing: there may be no fact of the matter whether a particular Carnap-style conditional for "water" or "pain" is analytic in English, or in some sub-community of English. In addition, Lewis's account is neutral on the existence and nature of "concepts," understood as psychological entities. It is not assumed that our concept of pain, for example, is in some sense decomposable into more basic concepts. Finally, it is not assumed that predicates like "pain" can be analyzed by a simple list of individually necessary and jointly sufficient predicates. The analysis takes the form of a rather complicated Carnap conditional.

32.3 Contingent Identity

Suppose the folk theory for pain looks something like this:

(1) Pain is typically caused by injuries, it tends to cause distress, etc.

Suppose further that physiological investigations reveal that the role characterized by (1) is realized by C-fiber firing:

(2) C-fiber firing is (the only state that is) typically caused by injuries, tends to cause distress, and so on.

(1) and (2) logically entail that pain is C-fiber firing. More generally, as long as the role characterized by (1) is uniquely realized by some brain state or other, it follows that pain is identical to that brain state. This is Lewis's "argument for the identity theory" (1966).

Lewis is confident that some premise along the lines of (2) is true, although this is of course empirical speculation: for all we know in the armchair, it could turn out that the role of pain is occupied by non-physical perturbations of ectoplasma. Premise (1) is also subject to empirical tests: it might turn out that nothing occupies the folk psychological role of pain – even if the role is weakened in the way I suggested in the previous section. In order to separate the empirical and non-empirical premises, it may be advisable to replace premise (1) by the corresponding Carnap conditional:

(1′) *If* some state is typically caused by injuries etc., *then* pain is typically caused by injuries etc.

(1′) and (2) still logically entail (3), but (1′) no longer contains the empirical assumption that something plays the pain role.

One might be suspicious about the parenthetical uniqueness clause in (2). Science can tell us that C-fiber firing plays the pain role, but can it also tell us that the role is not also played by something else, something non-physical (see Block and Stalnaker 1999)? The answer depends on the details of the role. If part of a role is that *x* is the only thing that does so-and-so, then it can't turn out that two different things fully realize the role. But Lewis is anyway not committed to the view that premise (2) can be conclusively established by science. The reasons for believing (2) may include considerations of theoretical simplicity or parsimony. What's important for Lewis is that the totality of physical truths entails the psychological truths, not that we can actually derive the psychological truths from our present knowledge about physics. In this context, the totality of physical truths must be understood as including a "that's all" clause, since otherwise all kinds of negative truths will be left out. In this sense, the physical truths do rule out that the pain role is realized by some non-physical state along with C-fiber firing.

On the other hand, it might turn out that no unique physical state occupies the pain role (even if we focus on a single species or individual). For one thing, there are many kinds of pain: toothache, headache, etc. What if these correspond to very different biological states? Even a psychologically determinate type of pain might involve different neural mechanisms on different occasions. There could also be more or less inclusive ways of identifying a realizer: perhaps C-fiber firing is an equally good candidate for playing the pain role in a given individual as some much wider brain state that includes the C-fiber firing. In these cases, Lewis's account would probably say either that "pain" is indeterminate between different candidates or that it does denote a state that is physically and biologically rather disjunctive (compare 1994, 305).

In Lewis's argument for the identity theory, the identity of mental states with biological states follows logically from folk-psychological definitions and broadly physical facts. There are no gaps to be filled by further considerations of simplicity and parsimony. There is also no logical room for *role functionalism*: the view that pain isn't the realizer of the pain role, but the higher-level property of being in some state or other that realizes the role. According to Lewis, this flatly contradicts the folk-psychological characterization of pain. If pain is defined as the state that does so-and-so, and C-fiber firing is the state that does so-and-so, then we aren't free to say that pain isn't actually the state that does so-and-so, but rather the property of being in some state or other that does so-and-so (see 1994, 307f.).

This means that Lewis faces the stock objection to the identity theory: if pain is C-fiber firing, then only creatures with C-fibers can have pain; but the folk understanding of psychological terms surely doesn't rule out that creatures of radically different kinds can have pain. Lewis offers two replies. First, he suggests to distinguish *pain* from *having pain*: while pain is defined by its causal-functional profile and must therefore be identified with the realizer, having pain is the higher-level property that is common among all creatures whose state occupies the role of pain in their respective cognitive architecture (see Lewis 1966, 101f.; 1994, 307). Lewis's second reply is that the identity of pain with a particular realizer state is contingent and kind-relative: *in humans, at the actual world*, pain is C-fiber firing (or whatever); in other creatures and at other worlds, pain may be something else (see 1969, 25; 1980a; 1983b, 43–5; 1986b, 267f.; 1994, 305–8); this position is also defended in Braddon-Mitchell and Jackson (1996), which is generally an excellent introduction to a theory of mind very close to Lewis's.

How can an identity be relative to a kind or world? First of all, whether or not something realizes a theoretical role can depend on a world, a time or other factors: H_2O satisfies "*x* covers a large part of the Earth" at the actual world, but not at other possible worlds; Barack Obama satisfies "*x* is president of the US" in 2012, but not in 2021; Lake Burley Griffin satisfies "*x* is the closest lake" relative

to my present location, but not relative to other locations. When a term is implicitly defined by a matrix like this, we have a choice of either fixing the relevant parameters once and for all or letting the term inherit the referential shiftiness of the matrix. Definite descriptions are usually shifty: "the closest lake" refers to different things in different contexts and under different embeddings. According to Lewis, mental terms are equally shifty: "pain" behaves just like "the state that plays the pain role." What this picks out depends on the contextually salient type of creature; in humans, the state that plays the pain role may be C-fiber firing, in Martians it may be something else. Thus "pain" denotes different states in different contexts and under different embeddings.

I find it rather implausible that expressions like "pain" behave in this manner. In my view, the semantically more plausible choice is to say that mental terms rigidly denote the relevant higher-level property. But then what about Lewis's argument against this proposal? It is true that "pain" must denote the realizer of the pain role. But the higher-level property may actually qualify as a realizer. This is obscured by formulations like (1), in which "pain" appears to pick out a particular node in a causal structure, but perhaps this is not the best regimentation of folk psychology. When a sharp pain causes me to withdraw my hand, then what does the causing is arguably not the universal pain, but a concrete *occurrence* or *instantiation* of pain. Suppose we rewrite the folk psychological definition along these lines, characterizing pain as a property x such that occurrences of x are typically caused by injuries, cause distress, etc. Now whenever we are in pain, we instantiate the higher-level property of being in some state or other with such-and-such typical causes and effects in creatures of our type. What brought it about that we instantiate this property? Often the cause might be an injury. And often instantiations of this property lie causally upstream of various signs of distress. So the higher-level property can realize the rewritten pain role.

Lewis notes this possibility in (1994, 307), but objects that the higher-level property is too disjunctive, "and therefore no events are essentially havings of it." He also complains that admitting both the lower-level and the higher-level property as causally efficacious "would lead to absurd double-counting of causes."

But who said that the events that are caused by injuries must be *essentially* havings of the property x? The modal individuation of events is notoriously murky, and the folk can hardly be assumed to have a settled opinion on this matter (see Bennett 1988). Moreover, if a particular pain event is contingently an occurrence of the higher-level property, then the very same event can also be an occurrence of a lower-level property. There is no double counting. This is exactly what Lewis says in 1997b (341f.) in response to the closely related question of whether dispositional properties or their categorical bases should be regarded as causally efficacious. "The very same event," he writes, "that is essentially a having of some causal basis of a certain disposition is also accidentally a having of the disposition itself. So an effect of this event is caused by a having of the basis, and caused also by a having of the disposition. But since these havings are one and the same event, there is no redundant causation." (Compare also Lewis 1986a, 223f.; 1997a, 142–4) I see no reason why the same can't be said for functional properties like pain and their lower-level bases.

One might fear that this solution is overly permissive. To use an example of Lewis's, suppose Mary dies because she put an aluminium ladder against a power line. We want to say that the ladder's electrical conductivity causally contributed to Mary's death, but not the ladder's opacity. However, the causal basis of electrical conductivity in aluminium is the same as the basis of its opacity. So the event that caused Mary's death is just as much an occurrence of conductivity as an occurrence of opacity. We somehow want to say that the event caused the death only in one of its two guises: Mary died because her ladder conducts electricity, not because it is opaque. One difference that seems relevant here is emphasized in Jackson and Pettit (1990): Mary's death is counterfactually invariant under substantial variations of its cause as long as the ladder's electrical conductivity is held fixed.

Thus, if Mary had used a different ladder, made of different material, she would still have died, as long as her ladder conducts electricity. By contrast, she wouldn't have died if she had used an equally opaque ladder made of wood. These considerations also explain why causal explanations in terms of higher-level, dispositional, or functional properties are often better than explanations in terms of their lower-level bases: the higher-level explanation says less about how the effect actually came about, but more about what would have happened under counterfactual circumstances. This is in part why psychological explanations are often more useful than explanations in terms of underlying neurobiological events.

Jackson and Pettit go on to suggest that higher-level properties are not involved in *real* causation at all, because all the causal work is done by more lower-level properties. This is reminiscent of Jaegwon Kim's "causal exclusion argument" against the efficacy of functional properties (see e.g. Kim 1998). Such arguments generally rely on a strongly anti-Humean and anti-Lewisian account of causation as a special kind of "production." If causation is something like counterfactual dependence or influence, then it isn't hard to see how higher-level properties can be involved in genuine causation. If Mary hadn't used a ladder that conducts electricity, she wouldn't have died. (It may also be worth pointing out that even if one believes in an anti-Humean force of production, one should accept that not all causation is by production, given the pervasiveness of double prevention, see Schaffer 2000; Hall 2004.)

Now I suggested that it isn't really properties that cause, but instantiations of properties. Our problem was that in Mary's case, there is a single event *C* which is an instantiation both of opacity and of electrical conductivity. Following Lewis (2004a), let's say that *C* is a cause of Mary's death *D* iff counterfactual variations of *C* go along with variations of *D*. To get the desired outcome, we might then suggest that in a context where *E* is described as an instantiation of electrical conductivity, the relevant counterfactual variations are variations with respect to conductivity. (See Lewis 2004b on the context-dependence of our individuation of events across worlds.) This might explain why Mary's death is caused by the instantiation of conductivity and not by the instantiation of opacity, although these instantiations are one and the same.

With role functionalism as a genuine option, the terminology gets a bit confusing, because we now have two different "roles" with correspondingly different "realizers." First of all, there are the folk-psychological roles for states like pain, expressed by a matrix like "occurrences of *x* are typically caused by injuries, cause distress," etc. This matrix expresses a second-order property, a property of properties. On the present proposal, the matrix is realized by the first-order property of being in some state or other occurrences of which are (in creatures of the salient kind) typically caused by injuries, cause distress, etc. This is a first-order property because it applies to individuals rather than properties. It is nevertheless higher-level in the sense that it is largely neutral on the physical or biological constitution of the relevant individuals. The higher-level property is a realizer of the pain role. On the other hand, relative to a particular type of individual, this higher-level property determines another role – another second-order property: to be in pain is to be in some state *y* that typically does such-and-such in creatures of the relevant kind. For humans, the state that does such-and-such might be C-fiber firing. One might therefore say that C-fiber firing is a "realizer of the pain role," but this pain role is not the role that defines "pain."

It may seem odd that the realizer of the pain role (the one that defines "pain") can effectively be read off from the role itself. We don't have to wait for science to find out that pain is the property of being in some state or other that does such-and-such. But this happens quite often in Canberra planning. Consider the role expressed by "*x* is a property which applies to all and only unmarried adult men," which we might have retrieved from a somewhat simplistic folk theory of bachelorhood. Again, this is a second-order condition, and we can immediately name a property that satisfies the matrix: being an unmarried adult man.

Does it matter whether psychological properties are identical to higher-level functional properties or to lower-level biological properties? Both accounts can agree on what there is, and even on the truth-conditions of sentences like "Fred has pain." Their disagreement only concerns the reference of singular terms like "pain." In his later writings, Lewis therefore suggests that the disagreement is superficial and unimportant (see e.g. 1994, 307; 1997a, 142–4; 2004b, 281). On the other hand, his identification of psychological states with brain states creates follow-up problems that may not arise on the alternative proposal. For example, according to Lewis, folk psychology says that people who are in pain typically know what state they are in: their evidence rules out any possibility where they are not in pain. If pain is C-fiber firing, this would mean that their evidence rules out every possibility where their C-fibers aren't firing. That seems false. Lewis concludes that this part of folk psychology must be rejected (see 1995, 327f.). If pain is a higher-level property, then knowledge that you are in pain is not knowledge that you are in a certain physiological state. Rather, it is knowledge about the functional profile of your current state. And it is not too implausible that when you introspectively recognize a state as pain, then you recognize it, for example, as a state that people are generally inclined to avoid (see Armstrong 1968, 96–9; see also Stoljar, chapter 33, this volume).

32.4 Beliefs, Desires, Decisions

A central part of folk psychology concerns the interaction of beliefs, desires, and choices. Crudely put, people typically do what they believe will bring about what they desire. Why does Mary play the cello at 3:00 am? Because she wants to annoy her neighbor and believes that her musical performance is a good way to achieve that. Other facts about Mary's attitudes are also involved, although we would rarely bother to mention them: Mary's desire to annoy her neighbor is not trumped by other desires, she believes that her neighbor will hear the cello, that he won't respond by setting fire to her apartment, and so on. Mary's behavior is explained not by a single belief and desire, but by a whole system of beliefs, desires, and possibly further attitudes.

There is an infinite number of possible systems of attitudes. Not only are there infinitely many things a person could in principle believe and desire, these attitudes also come in many different degrees: Mary can be more or less certain about how her neighbor will react, and her desire to annoy him may be stronger or weaker. According to Lewis, this part of folk psychology, when systematized, "should look a lot like Bayesian decision theory" (1979, 149). Bayesian decision theory represents a system of beliefs and desires by a pair of a probability function P and a utility function U. Both functions assign numbers to propositions. The P value assigned to a proposition represents the agent's degree of confidence that the proposition is true. (It is not assumed that the agent knows the proposition's objective probability.) The U value represents the degree to which she would like it to be the case that the proposition is true. (It is not assumed that this is a matter of material well-being.) Given a choice between some actions, decision theory then says that the agent, if rational, makes true whatever option A has the highest expected utility, defined as

$$EU(A) = \sum_{S \in W} P(S)U(S \,\&\, A),$$

where W is a suitable partition of propositions. What exactly makes a partition "suitable" is a controversial matter. Lewis (1981a, 11) suggests that each member of W should be a "maximally specific proposition about how the things [the agent] cares about do and do not depend on his present actions."

Decision theory thus describes a simple connection between any system of (coherent) beliefs and desires and a corresponding set of choice dispositions. Hence one can to some extent read off a rational agent's attitudes from the choices she is disposed to make when confronted with a given set of options. This approach was already developed by Ramsey (1931c), who also proved a *representation theorem* showing that if an agent's choice dispositions satisfy certain qualitative constraints, then there is *unique* system of beliefs and desires that matches her dispositions. (For utilities, "unique" means unique up to positive affine transformation, since utility scales have arbitrary unit and zero.)

More streamlined results to the same effect have since been established, for example by Savage (1954). However, the qualitative constraints assumed in these theorems are extremely strong. On Lewis's account, this is somewhat ameliorated by the fact that the agent's actual choice dispositions are less important than the typical dispositions of other (perhaps merely possible) agents in the same state (see Lewis 1974, 119–21; 1994, 321f., 324 fn. 42). Nevertheless, Lewis clearly deemed the constraints too strong: choice dispositions, he says, do not fully determine an agent's system of beliefs and desires. Further constraints must be invoked (see 1983b, 50–2; 1986c, 107f.).

Some of these further constraints concern the way systems of belief and desire typically change under the impact of perceptual stimulation. Folk psychology says that under normal circumstances, people who are falling down a crevasse realize that they are falling; that is, they come to assign high probability to the hypothesis that they are falling. Similarly, when people with functioning eye sight confront a red wall, they typically come to believe that there is something red in front of them (see Lewis 1979, 514, 534; 1980b, 274; 1983a, 380; 1983b, 50; 1994, 299f.; 1997b).

Yet further (and evidently non-causal) compartments of folk psychology constrain the kinds of things for which people have non-instrumental desires, and the kinds of hypotheses they find *a priori* plausible or implausible (see Lewis 1974, 112f.; 1986c, 38f., 107; 1994, 320). Lewis suggests that objective naturalness might play a role here, but it remains unclear what exactly this amounts to (see 1983b, 52–4; 1986c, 38f., 107; 1994, 320).

The total picture then looks as follows. Folk psychology assigns a complex role to entire systems of belief and desire, as represented by a pair of a probability function and a utility function. The characteristic role of such a system is, in the first place, to cause certain kinds of behavior under certain conditions. In addition, folk psychology says that systems of beliefs and desires tend to change in a certain way in response to perceptual input, and that they tend to satisfy some general constraints of rationality. If a brain state comes sufficiently close to playing this role, then it can be identified with the relevant system of beliefs and desires.

One can think of folk psychology as a set of interpretation rules: if a state typically gives rise to such-and-such behavioral dispositions, has such-and-such typical causes, etc., then it can be interpreted as being a system of belief and desire with such-and-such content. But it would be wrong to conclude that for Lewis, an agent's beliefs and desires are somehow dependent on, or relative to, an external interpreter. The rules of folk psychology define *what it is* to have certain beliefs and desires. Whether an agent satisfies the definition is then a perfectly objective matter. An agent's beliefs and desires are no less objective and real than her mass and height.

A few aspects of Lewis's account may deserve special emphasis. One is the built-in holism about beliefs and desires. Decision theory, as outlined above, says nothing about individual beliefs and desires. This may be one reason for Lewis's suggestion that "beliefs" may be a "bogus plural" (1994, 311). According to Lewis, it is an open empirical question whether our brains store information holistically, like a map or a connectionist network, or in discrete units. In the latter case, he argues, one might regard the individual units as individual beliefs. A more natural suggestion, I think, is to identify the property of believing a particular proposition p with (roughly) the property of having a system of beliefs and desires which assigns sufficiently high probability to p. A belief that p would then be an instantiation of this property. Following the remarks in the previous section,

one could even say that such individual beliefs can occupy causal roles: Mary's belief that her neighbor would get annoyed if she played the cello is a state which, on the background of such-and-such other states (and absences), causes her to play the cello. Unlike, for example, Mary's belief that Napoleon died on St. Helena, the belief about her neighbor is a causal difference-maker for Mary's behavior.

Many naturalistic accounts of mental content assume not only that there are individual beliefs, but that they really are stored in discrete units, perhaps further decomposed into individual "concepts." These concepts are then said to have their content in virtue of being normally "activated" in the presence of horses or other external objects. Lewis complains that this not only relies on contentious empirical assumptions about the architecture of the brain, but also largely ignores the folk psychological role of mental content (see 1994, 310–24). Notice that from a decision-theoretic perspective, what needs to be naturalized is not so much the content of beliefs and desires but their strength. That's because what primarily varies from agent to agent, and thus what needs to be explained by physical or functional differences between agents, are the degrees of belief and desire associated with any given proposition. I am not aware of any even remotely plausible answer to this question in terms of causal origins (or conscious phenomenology, for that matter).

A second fact I want to highlight is that on the Lewisian (or Ramseyan) account, the norms of decision theory are not so much normative or descriptive but *constitutive* of agents with beliefs and desires. It is often claimed that ordinary people systematically violate the norms of decision theory, for instance by cooperating in a prisoner dilemma or by rejecting offers in the ultimatum game. Such claims often rest on overly simplistic assumptions about the agents' beliefs and desires, such as an identification of utilities with monetary payoff. (For discussion, see e.g. Blackburn 1998 and Joyce 1999, chapter 2, but note that while Blackburn endorses the Ramseyean account, Joyce does not.) If an agent's beliefs and desires are *defined* in part as whatever probabilities and utilities make their choices come out rational, then it not easy to establish that the choices people make are generally not rational by the light of their beliefs and desire.

Nevertheless, standard decision theory includes idealizations that aren't part of folk psychology. In particular, it leaves little room for reasoning and *a priori* inquiry: by the axioms of the probability calculus, the tautologous proposition which is true at every possible world has probability 1, and whenever a proposition A has probability x, then any proposition entailed by A has probability greater than or equal to x. Contrary to popular belief, this does not have the absurd consequence that everyone should be certain that Hesperus is Phosphorus, or that Fermat's Last Theorem is true. This would only follows if one were to read "x is certain that S" as saying that the relevant agent assigns high probability to the proposition expressed by S. But there are good reasons to reject this hypothesis about attitude reports in English. The real "problem of logical omniscience" is that whatever belief is attributed with expressions like "x believes that Fermat's Last Theorem is true," it looks like the relevant proposition (for example about the meaning of certain mathematical symbols) is entailed by other propositions which the agent knows. Solving this problem arguably requires going beyond the assumption that beliefs can be modeled by a standard probability distribution. (See Lewis 1982, 103; Stalnaker 1984, 1991, 1999b for some attempts in this direction.)

A third and final aspect of Lewis's account that I want to mention is his insistence that mental content is narrow: that it never differs between intrinsic duplicates within the same world. This is partly due to his emphasis on the connection to behavioral dispositions. In Putnam's (1975) Twin Earth scenario, Oscar and his Twin Earth counterpart Twoscar are causally connected to chemically different substances, but they are disposed to display the exact same responses when put in the same situations. Moreover, on Lewis's account, an agent's attitudes are determined not only by the actual causes and effects of their inner state, but also by the role this state is disposed to play under other actual or hypothetical circumstances. The fact that on our planet, a certain type of belief state is

usually caused by the presence of H_2O therefore doesn't entail that the state's content somehow involves H_2O. After all, the very same state is usually caused by the presence of XYZ on Twin Earth.

What is revealed by the Twin Earth thought experiment (as well as, for example, Burge's (1979) arthritis example) is that ordinary-language attitude reports can be sensitive to differences in the agent's environment: at least in some contexts, one can truly say that Oscar believes that water covers a large part of his planet, while Twoscar does not. Lewis's explanation is that *de re* belief statements, of the form "*x* believes that *y* is *F*," mean that *y* satisfies some condition *G* such that *x* assigns high probability to the proposition that the *G* is *F*. Oscar, for example, assigns high probability to the proposition that the watery stuff in his surroundings covers a large part of his planet; this stuff is in fact water, that is, H_2O; hence we can truly say that Oscar believes that water covers a large part of his planet (see Lewis 1979, §13; 1981b, 412–14; 1986c, 32–4; 1994, 318f.; Cresswell and von Stechow 1982). The condition *G* captures the way the relevant object is "presented" to the agent, which often makes a difference to the agent's behavior (see Lewis 1983b, 50; 1979, 142f.; 1981b; 1994, 323f.).

The fact that Lewis attributes the same beliefs to Oscar and Twoscar does not mean that the content of their belief state is determined without reference to the external world. (Stalnaker 2004) draws this conclusion and concludes that Lewis must have endorsed a kind of conceptual role account on which the content of a belief state is fixed by its syntactical structure together with a naturalness constraint on its interpretation. But that was not Lewis's view. As we have seen, for Lewis, what determines the content of a system of beliefs and desires is not its inner structure, but its typical perceptual inputs and behavioral outputs.

32.5 Phenomenal Character

Many mental states have a distinctive qualitative or phenomenal character: pain has typical causes and effects, but it also has a typical feel. Lewis argues, plausibly enough, that it is not a contingent empirical discovery that pain feels painful: if something doesn't (typically) have the phenomenal character of pain, then it isn't pain. If pain can be analyzed by its functional profile, it follows that the same is true for the phenomenal character of pain. Indeed, according to Lewis, the phenomenal character of pain is simply the property which is satisfied by a state iff the state plays the pain role. But is the phenomenal character of pain really determined by the state's functional profile? Doesn't information about a state's causal or functional role leave it open how that state feels to its subject – and whether it feels like anything at all? Can't one imagine "zombies" that are physically and functionally just like us but lack phenomenal consciousness? Thus goes the conceivability argument against analytic functionalism (see Chalmers 1996, 93–171).

It is important to get the argument right. Analytic functionalism does not rule out the possibility that there are creatures physically just like us but without mental states. Suppose it turns out that non-physical perturbations of ectoplasma play the mental roles – which is perfectly compatible with analytic functionalism. Duplicating our physical bodies without duplicating the ectoplasma would then leave out the mental states. Chalmers therefore defines zombie worlds not as worlds physically like the actual world but without consciousness, but as worlds where $P \wedge \neg Q$ is true, where P is the totality of the actual physical truths and Q is (say) the claim that someone experiences pain. But again, a friend of analytic functionalism need not deny that there are worlds where $P \wedge \neg Q$ is true. If non-physical states turn out play the role of mental states, and the terms in Q rigidly denote the realizers, then $P \wedge \neg Q$ describes a genuine possibility. What analytic functionalism, combined with physicalism, has to deny is not the conceivability of a zombie world, nor the conceivability of there being a $P \wedge \neg Q$ world, but the conceivability of $P \wedge \neg Q$ itself, as a hypothesis about the actual world.

Since P includes all truths about the typical functional roles of our brain states, $P \wedge \neg Q$ entails something like the following.

> In creatures like us, there is a (unique) state that is typically caused by injuries, causes distress and a desire for the state to go away etc., but people who are in this state never experience pain.

According to analytic functionalism, this – understood as a hypothesis about our actual situation – is subtly incoherent. It is certainly hard to imagine how we could find out that the claim is true, or how we could ever *have* found that out. For my part, I think there is a reading on which the hypothesis is indeed coherent. But the issue is much less straightforward than often portrayed.

Perhaps a better argument in support of the same anti-physicalist conclusion looks at the information we receive when we have an experience. Suppose one morning you feel an unusual twitch in your leg. What do you learn when you notice this sensation? Unless you happen to be a neuroscientist, you presumably don't learn that you are in such-and-such a physiological state. If you've never had that sensation before and don't know what it is called, you arguably also don't learn that you're in a state with such-and-such typical causes and effects. What then is the information you acquire? It looks like what you learn is not, or not just, that you have certain physical or functional properties (including indexical properties). But then you seem to have information which is not analytically entailed by the totality of all physical truths.

This line of thought is related to Frank Jackson's (1982) knowledge argument, but it doesn't concern a peculiar situation in which someone knows all relevant physical facts. Moreover, I explicitly asked what information you acquire, not what you learn or come to know, nor how to understand "knowing what it's like." Lewis suggests that when Jackson's Mary comes to know what it's like to see red, she primarily acquires a new set of abilities (as well as indexical information and perhaps new forms of representation, see Lewis 1983c, 131f.; 1988, 268ff., 278f., 287, 290; 1994, 294): Mary learns to visually recognize and classify colors, to imagine red triangles, etc. (Lewis 1983c, 131; 1988, 285–8; 1995, 326f.). She can't learn anything else, because she already knows all the physical facts. By contrast, when you notice your twitch, you are obviously unaware of many physical facts. The problem is that none of these seem to be good candidates for the information you acquire, but it is highly implausible to say that you don't acquire any information at all.

Acknowledgments

Thanks to Jochen Faseler, Alan Hájek, Daniel Nolan, Adam Pautz, and Daniel Stoljar for helpful comment on an earlier version.

References

Armstrong, D.M. (1968). *A materialist theory of the mind*. London: Routledge.

Bennett, J. (1988). *Events and their names*. Oxford: Clarendon Press.

Blackburn, S. (1998). *Ruling passions: a theory of practical reasoning*. Oxford: Oxford University Press.

Block, N. & Stalnaker, R. (1999). Conceptual analysis, dualism, and the explanatory gap. *Philosophical Review*, 108, 1–46.

Braddon-Mitchell, D. and Jackson, F. (1996). *Philosophy of mind and cognition*. Oxford: Blackwell.

Burge, T. (1979). Individualism and the mental. In A.P. French, T.E. Uehling & H.K. Wettstein (Eds.), *Midwest studies in philosophy*, Vol. IV, pp. 73–121. Minneapolis: University of Minnesota Press.

Carnap, R. (1963). Replies and systematic exposition. In P.A. Schilpp (Ed.), *The philosophy of Rudolf Carnap*, pp. 859–1016. La Salle, IL: Open Court.
Chalmers, D. (1996). *The conscious mind*. New York: Oxford University Press.
Collins, J., Hall, N. & Paul, L.A. (Eds.) (2004). *Causation and counterfactuals*. Cambridge, MA: MIT Press.
Cresswell, M. & von Stechow, A. (1982). *De re* belief generalized. *Linguistics and Philosophy*, 5, 503–35.
Hall, N. (2004). Two mistakes about credence and chance. *Australasian Journal of Philosophy*, 82, 93–111.
Jackson, F. (1982). Epiphenomenal qualia. *Philosophical Quarterly*, 32, 127–36. Reprinted in Jackson (1998b).
Jackson, F. (1998a). *From metaphysics to ethics: a defence of conceptual analysis*. Oxford: Clarendon Press.
Jackson, F. (1998b). *Mind, method and conditionals: selected essays*. London: Routledge.
Jackson, F. & Pettit, P. (1990). Causation in the philosophy of mind. *Philosophy and Phenomenological Research Supplement*, 50, 195–214.
Joyce, J. (1999). *The foundations of causal decision theory*. Cambridge: Cambridge University Press.
Kim, J. (1998). *Mind in a physical world: an essay on the mind-body problem and mental causation*. Cambridge, MA: MIT Press.
Lewis, D. (1966). An argument for the identity theory. *Journal of Philosophy*, 63, 17–25. Reprinted with extensions in Lewis (1983c), pp. 99–107.
Lewis, D. (1969). Review of *Art, mind, and religion*. *Journal of Philosophy*, 66, 22–7.
Lewis, D. (1970). How to define theoretical terms. *Journal of Philosophy*, 67, 427–46. Reprinted in Lewis (1983c), pp. 78–95.
Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, 50, 249–58. Reprinted in Lewis (1999), pp. 248–61.
Lewis, D. (1974). Radical interpretation. *Synthese*, 23, 331–44. Reprinted in Lewis (1983c), pp. 108–18.
Lewis, D. (1979). Attitudes *de dicto* and *de se*. *Philosophical Review*, 88, 513–43. Reprinted in Lewis (1983c), pp. 133–56.
Lewis, D. (1980a). Mad pain and Martian pain. In N. Block (Ed.), *Readings in the philosophy of psychology*, Vol. I, Cambridge, MA: Harvard University Press, 216–22. Reprinted in Lewis (1983c), pp. 122–30.
Lewis, D. (1980b). Veridical hallucination and prosthetic vision. *Australasian Journal of Philosophy*, 58, 239–49. Reprinted in Lewis (1986d), pp. 273–86.
Lewis, D. (1981a). Causal decision theory. *Australasian Journal of Philosophy*, 59, 5–30. Reprinted in Lewis (1986d), pp. 305–37.
Lewis, D. (1981b). What puzzling Pierre does not believe. *Australasian Journal of Philosophy*, 59, 283–9. Reprinted in Lewis (1999), pp. 408–17.
Lewis, D. (1982). Logic for equivocators. *Noûs*, 16, 431–41.
Lewis, D. (1983a). Individuation by acquaintance and by stipulation. *Philosophical Review*, 92, 3–32. Reprinted in Lewis (1999), pp. 373–402.
Lewis, D. (1983b). New work for a theory of universals. *Australasian Journal of Philosophy*, 61, 343–77. Reprinted in Lewis (1999), pp. 8–55.
Lewis, D. (1983c). *Philosophical papers*, Vol. I. New York and Oxford: Oxford University Press.
Lewis, D. (1984). Putnam's paradox. *Australasian Journal of Philosophy*, 61, 343–77. Reprinted in Lewis (1999), pp. 56–77.
Lewis, D. (1986a). Causal explanation. In Lewis (1986d), pp. 214–40.
Lewis, D. (1986b). Events. In Lewis (1986d), pp. 241–69.
Lewis, D. (1986c). *On the plurality of worlds*. Oxford: Blackwell.
Lewis, D. (1986d). *Philosophical papers*, Vol. II. New York and Oxford: Oxford University Press.
Lewis, D. (1988). What experience teaches. *Proceedings of the Russellian Society*, 13, 29–57. Reprinted in Lewis (1999), pp. 262–90.
Lewis, D. (1989). Dispositional theories of value. *Proceedings of the Aristotelian Society*, Suppl. Vol. 63, 113–37. Reprinted in Lewis (2000), pp. 68–94.
Lewis, D. (1994). Reduction of mind. In S. Guttenplan (Ed.), *A companion to the philosophy of mind,* pp. 412–31. Oxford: Blackwell. Reprinted in Lewis (1999), pp. 291–324.
Lewis, D. (1995). Should a materialist believe in qualia? *Australasian Journal of Philosophy*, 73, 140–4. Reprinted in Lewis (1999), pp. 325–31.

Lewis, D. (1996). Desire as belief II. *Mind*, 105, 303–13. Reprinted in Lewis (2000), pp. 55–67.
Lewis, D. (1997a). Finkish dispositions. *Philosophical Quarterly*, 47, 143–58. Reprinted in Lewis (1999), pp. 133–51.
Lewis, D. (1997b). Naming the colours. *Australasian Journal of Philosophy*, 75, 325–42. Reprinted in Lewis (1999), pp. 332–58.
Lewis, D. (1999). *Papers in metaphysics and epistemology*. Cambridge: Cambridge University Press.
Lewis, D. (2000). *Papers in ethics and social philosophy*. Cambridge: Cambridge University Press.
Lewis, D. (2002). Tharp's third theorem. *Analysis*, 62, 95–7.
Lewis, D. (2004a). Causation as influence. In Collins *et al*. (2004), pp. 75–107.
Lewis, D. (2004b). Void and object. In Collins *et al*. (2004), pp. 277–91.
Lewis, D. (2009). Ramseyan humility. In D. Braddon-Mitchell & R. Nola (Eds.), *Conceptual analysis and philosophical naturalism*, pp. 203–22. Cambridge, MA: MIT Press.
Putnam, H. (1975). The meaning of "meaning." In K. Gunderson (Ed.), *Language, mind, and knowledge*, pp. 131–93. Minnesota: University of Minneapolis Press.
Ramsey, F. (1931a). *Foundations of mathematics and other essays*. London: Routledge & P. Kegan.
Ramsey, F. (1931b). Theories. In Ramsey (1931a).
Ramsey, F. (1931c). Truth and probability. In Ramsey (1931a).
Reichenbach, H. (1938). *Experience and prediction*. Chicago: University of Chicago Press.
Ryle, G. (1949). *The concept of mind*. Chicago: University of Chicago Press.
Savage, L. (1954). *The foundations of statistics*. New York: Wiley.
Schaffer, J. (2000). Causation by disconnection. *Philosophy of Science*, 67(2), 285–300.
Stalnaker, R. (1984). *Inquiry*. Cambridge, MA: MIT Press.
Stalnaker, R. (1991). The problem of logical omniscience I. *Synthese*, 89, 425–40. In Stalnaker (1999a).
Stalnaker, R. (1999a). *Context and content*. Oxford: Oxford University Press.
Stalnaker, R. (1999b). The problem of logical omniscience II. In Stalnaker (1999a), pp. 255–73.
Stalnaker, R. (2004). Lewis on intentionality. *Australasian Journal of Philosophy*, 82, 199–212.

33

Lewis on Materialism and Experience

DANIEL STOLJAR

33.1 Introduction

The traditional question in philosophy of mind of the compossibility or cotenability of materialism, on the one hand, and the existence of experience on the other is a theme throughout David Lewis's work. His first published paper, "An Argument for the Identity Theory," (1966) , was a defence of materialism. And one of his most famous posthumous papers, "Ramseyan Humility" (2009), contains a discussion of the identification thesis, the rejection of which is crucial to his account of experience. In between, papers such as "Theoretical and Psychophysical Identifications" (1972), "Mad Pain and Martian Pain" (1978), "New Work for a Theory of Universals" (1983a), "What Experience Teaches" (1988), "Reduction of Mind" (1994), "Should a Materialist Believe in Qualia?" (1995), and others contain a systematic and extremely influential approach to this topic.

This approach might usefully be viewed as the conjunction of four elements. *Element #1* is materialism[1] for which Lewis gave a distinctive and well-known characterization. *Element #2* is an account of what experience is: experience is the occupant of a certain functional role implicitly set out by folk psychology. *Element #3* is an account of the source of the tension between experience and materialism: the truth of materialism is apparently inconsistent with the existence of experience if we adopt a conception according to which experience is construed as satisfying all of the requirements of folk psychology. *Element #4* is a strategy for resolving the tension: Lewis argues that while the existence of experience may preclude materialism on the conception just mentioned, it is possible to articulate a different conception, according to which experience is construed as satisfying most (but not quite all) of the requirements of folk psychology. In this sense "it is not altogether wrong to call him an eliminativist, but see how little he eliminates and how much he retains" (1995, p. 329).

In the first part of this paper I will review these four elements of Lewis's account of materialism and experience. In the second I will consider two recent objections to that account. The first argues that knowledge-how is a certain kind of knowledge-that and in consequence Lewis's well-known "ability hypothesis" fails. The second argues that if Lewis's contextualist approach to epistemology is correct, his rejection of the identification thesis is impossible. I will suggest that Lewis has the

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

resources to answer both objections, but I will end by stating where in my view the real problems for Lewis lie.

33.2 Element #1: Materialism

Lewis did not just give *a* distinctive and well-known characterization of materialism, he gave *two*: one in terms of fundamental properties, and one in terms of supervenience. The fundamental properties definition starts, he says, from something he believes *a priori* (1994, 291), namely, that any possible world at all instantiates a relatively small class of fundamental properties, where "fundamental" or "perfectly natural" properties are (among other things) "not at all disjunctive, or determinable, or negative. They render their instances perfectly similar in some respect. They are intrinsic; and all other intrinsic properties supervene on them" (2009, 204). On the assumption that what Lewis believes here is not merely *a priori* but necessary, one might think that materialism is the thesis that every fundamental property in any possible world at all is physical. But this construes materialism as necessary if true, when it is usually taken to be contingent.

A better approach is to provide what is sometimes called a "world-relative" definition of materialism, that is, one that tells us what it is for materialism to be true at some possible world arbitrarily chosen:

> **M1.** Materialism is true at a possible world w iff every fundamental property instantiated at w is physical.

Given this definition, the materialist *about a possible world* W believes that every fundamental property at W is physical. But of course we don't normally speak of philosophers as materialist (or not) about this or that possible world; we speak of them as materialists (or not) without qualification, that is, as just materialists. What is it then according to M1 to be a materialist without qualification? To be a materialist without qualification is from this point of view to be materialist about one possible world *in particular*, namely, the actual world. This has the benefit of portraying materialism as if true a contingent truth, because it is a contingent truth (if it is a truth) that the fundamental properties instantiated at the actual world are physical properties. The dualist or the vitalist, for example, will say that at least some of these properties are psychological or biological. Neither dualism nor vitalism is impossible, it is simply that neither is true, or at any rate not if materialism is true.

Turning to the supervenience definition, this focuses on the idea of two possible worlds being, as Lewis says, duplicates of one another (1983a, 27). Lewis starts his discussion here by portraying materialism as the thesis that, for any two possible worlds, if they are duplicates with respect to the physical, they are duplicates *simpliciter*. But, as he notes, this again wrongly treats materialism as necessary if true. As before, a world-relative account is better:

> **M2.** Materialism is true at a possible world w iff for any possible world w* if w and w* are physical duplicates, they are duplicates *simpliciter*.

Like M1, M2 permits materialism to be a contingent truth, if it is true at all. If dualism is true at the actual world, there will be at a least one world that is a physical duplicate of the actual world but is not a duplicate *simpliciter*, that is, because it is different psychologically.

As Lewis notes, M2 faces an important counterexample. Consider a putatively possible world ε, which is physically exactly like our world but which contains some additional epiphenomenal ectoplasm: some stuff that is non-physical and yet does not in anyway disrupt the physical goings on in

the world. Such a world seems possible; in particular, the materialist should not deny its possibility in view of the contingency of materialism. But M2 has the consequence that if materialism is true, ε is impossible. For if materialism is true, and M2 is the definition of materialism, any world physically exactly like the actual world is exactly like it in all respects. But ε is a counterexample to this: it is a world that is physically like the actual world but which is not exactly like it. Hence, M2 is mistaken.

To deal with this problem Lewis appeals to the notion of an alien property, where a "property is alien to a world iff (1) it is not instantiated by an inhabitant of that world; and (2) it is not analysable as a conjunction of, or as a structural property constructed out of, natural properties all of which are instantiated by inhabitants of that world" (1983a, 37).[2] In the light of this, it is possible to modify M2 as follows:

> **M3.** Materialism is true at w iff for any possible world w*, if (a) w* and w are physical duplicates, and (b) w* contains no natural properties alien to w; then w* and w are duplicates *simpliciter*.

Unlike M2, M3 does not have the truth of materialism entailing that ε is impossible, since ε is a world that includes properties alien to the actual world, at least if we suppose that epiphenomenal ectoplasm is or instantiates alien properties.

What is the difference between the fundamental property and supervenience definitions, that is, between M3 and M1? One might think they are different in that one makes use of the notion of a fundamental property while the other does not. But this is not so. First, fundamental or at least natural properties are implicit in Lewis's definition of duplication and so in his definition of supervenience (Lewis 1983a, 27). Second, as we have seen, Lewis appeals to alien properties in M3, and these are defined in terms of natural properties.

One might think that the two definitions are different in that one makes use of modal notions and in particular supervenience while the other does not. But this is not so either. Fundamental properties, as Lewis conceives them, are properties "which figure in a minimal basis on which all else supervenes" (2009, 205), so there is clearly a modal element in M1 too. Indeed, both M1 and M3 include at some level both supervenience and fundamentality.

One might also think that the two definitions are equivalent or at least that Lewis thinks that they are. But Lewis does not say that they are equivalent. In fact, in "New Work for a Theory of Universals," the 1983 paper in which he explicitly advances the supervenience definition, he considers the fundamental properties definition[3] and rejects it in the following passage:

> Couldn't there be a natural property X (in the nature of the case, it is hard to name an example!) which is shared by the physical brains in worlds like ours and the immaterial spirits that inhabit other worlds? Or by this worldly quarks and certain otherworldly particles that cannot exist under our physics? Physics could quite properly make no mention of a natural property of this sort. It is enough to recognize the special case applicable to our world, X-cum-physicality, brainhood or quarkhood as it might be. Then, if by physical properties we mean those that are mentioned in the language of physics, a Materialist ought not to hold that all natural properties are instantiated in our world are physical properties. (1983a, 34)

The suggestion here seems to be that the fundamental properties definition – M1, in our terms – is mistaken because it erroneously requires materialism to be false in the situation in which X is instantiated. After making this suggestion, Lewis goes on immediately to consider the supervenience definition (i.e. M3 in our terms), which suggests that this is his reason for abandoning M1 in favor of M3.

But it is very difficult to see this passage as containing a good reason for abandoning M1 in favour of M3. First, and this is merely ad hominem, in other papers (e.g. 1994) Lewis operates with the fundamental properties formulation and there is no suggestion that it is open to an objection that

the supervenience definition is not. Second, if it worked against M1, this objection would also work against M3. For suppose Lewis's property X is instantiated. Presumably X is either necessitated by the physical or it is not. Suppose it is not necessitated; then X presents a problem for M3 as well as M1. For in that case there is a world which (a) is a physical duplicate of our world; (b) instantiates no further alien properties; but (c) is not a duplicate *simpliciter*, that is, because it does not instantiate X. Now suppose X is necessitated by the physical; then X presents no problem for M3 but likewise it presents no problem for M1. For a proponent of M1 will reply that X is a property like many others which is necessitated by fundamental physical properties but which is not itself a fundamental physical property. Finally, it is not clear that the objection *does* work. For suppose there were a fundamental property unrecognized by physics but which imposed a genuine objective similarity on physical brains and immaterial spirits. It is quite natural to think that in this case that materialism *is* false, just as both M1 and M3 predict; hence there is no problem for either here.

Part of what makes this objection difficult is the reference in it to Lewis's account of what it is for a property to be a physical property. No definition of materialism, and in particular none of M1–M3, is complete without an account of what a physical property is. As Lewis says in the passage just quoted, "by physical properties we mean those mentioned in the language of physics," and in the same paper he says which physics he has in mind: something very like the one used by contemporary, actual, physicists, though "presumably somewhat improved" (1983a, 33–4). But such an account faces serious problems. For one thing, could there not be properties that are both physical *and* alien, properties instantiated at worlds whose physics is not definable in terms of ours, and yet is in other respects recognizably like ours?[4] Moreover, consider what Lewis in a later paper (2009, 205) calls *idlers*, that is, fundamental properties which are instantiated at our world but which play no role in physical, or indeed any, theory. If there are such properties, then no matter what the intrinsic nature of such properties, it immediately follows from M1 and M3, together with Lewis's account of a physical property, that materialism is false, for then there will be fundamental properties that are not physical by his definition. Lewis argues that there is no reason to deny that idlers are instantiated (2009, 214). But then by his own account of physical properties he has no reason to believe physicalism!

In view of the fact that is unclear both what the relation is between M1 and M3, and what account of a physical property they presuppose, there are clearly some open questions for Lewis's approach to materialism.

33.3 Element #2: Experience

In ordinary English, an experience is an event of some sort that in a certain hard-to-specify way engages our (or some salient subject's) psychology in certain distinctive ways. Take the royal wedding between Prince Dullsville and Princess Whatever-her-name-is. That was an experience, or at least was so for the 2 billion people that watched it on television. The formation of the solar system, by contrast, was an event but not an experience because it did not have the right sort of engagement with our (or anyone's) psychology. Obviously spelling out the right sort of engagement is a difficult matter. But, however this is done, an experience need not in this ordinary sense be a psychological event like a judgement or the onset or persistence of pain. The royal wedding is not a psychological event though it is an experience. One can say, for example, "the experience left the princess a changed person," meaning that the wedding left her changed.

In the dialect of English spoken by philosophers of mind, "experience" is usually restricted to psychological events, and in particular to psychological events associated with sensory perception and bodily sensation, such as tasting the champagne, catching a glimpse of the abbey, feeling the

ring in one's pocket, and so on. In some parts of his writing, Lewis uses "experience" in this way: "I will say *experiences* when I mean particular events of experiencing" (1995, 326), and it is clear in the context that he has in mind psychological events. Lewis also talks of experiential states, which are types of events of experiencing, or perhaps properties of these events. The two ideas are not unrelated since an experiencing event is naturally thought of as consisting (at least in part) in the onset or having of an experiential state, and indeed the word "experience" might legitimately be used for both. But they are also different: an event is something that happens at a particular time and place and which cannot be repeated; an experiential state, by contrast, is a type of thing or a property, something that two people can have or be in, and one person at different times can have or be in.

If experience (in the restricted sense that operates in philosophy of mind) is an event that consists in the onset or the persistence of an experiential state, what is it that makes something an experiential state? For Lewis, experiential states are implicitly defined by folk psychology. Folk psychology is a tacitly known theory or body of information about how people in general act and think; so folk psychology might say that experiences have this sort of feature, and are typically brought about this way, and so on. Lewis thinks that if we can isolate the key principles or theses of folk psychology that concern experience, then in principle one could use them to construct an explicit definition of what experience is, using the famous Ramsey-Carnap technique (see Lewis 1972, 1994, 1997). We will not go into that technique here; for us the key point is that experience is that thing which satisfies all or most of the relevant principles of folk psychology.

What then are the relevant principles of folk psychology? Lewis did not give a detailed analysis of what folk psychology says about experience; at one point he says he offers "recipes for analysis" rather than fully fledged accounts (1994, 298). Nevertheless, it is possible to draw out from his writings some central principles, some of which are more controversial or general than others, some of which overlap, and some of which might turn out to be explained in terms of others. These are:

Causal Thesis: when you have an experience[5] you are in some inner state which typically causes you to behave in various ways, to form other mental states, and which is typically caused by certain things in the world. (Lewis 1966, 1972, 1978)

Informational Thesis: when you have an experience, you are in state with a certain informational content. (Lewis 1983c)

Belief Thesis: when you have an experience with the informational content p you are disposed to form a belief that p (or a similar belief). (Lewis 1983a, 1994)

Something It Is Like Thesis: when you have an experience, there is something it is like for you to have the experience. (Lewis 1988)

Knowing What It Is Like Thesis: when you have an experience, you know what it is like to have that experience. (Lewis 1988, 1995)

Ability Thesis: when you have an experience you have an ability to imagine the experience, to remember the experience, and to recognize the experience. (Lewis 1988, 1994, 1995)

Identification Thesis: when you have an experience you know the essence of the experience, that is, exactly what it is in an "uncommonly demanding sense." (Lewis 1995, 1997)

Self-Intimation Thesis: when you have an experience you know that you are having the experience. (Lewis 1972, 1996)[6]

Putting these theses together we arrive at the view that an experiential state is a state that satisfies all or most of them. As we will see, "all or most" part is important. Consider the view that, by definition, an experiential state satisfies all (and not merely most) of these. And suppose now we discover that nothing that exists in a physical world – that is, a world at which materialism is true – could

satisfy all of them. It would follow that nothing that exists in a physical world could be an experience; that is, materialism and the existence of experience are not cotenable. On the other hand, if we require only that an experiential state satisfies most of these this result may well be avoided.

33.4 Element #3: Materialism and Experience in Tension

We have considered what materialism is for Lewis, and what experience is; I now turn to the two main arguments he discusses against the cotenability of materialism and the existence of experience.

The first – the *identification argument*, I will call it – starts from an inconsistency between materialism, the identification thesis, and some agreed-on facts (see Lewis 1995). Consider the following claims:

(1) If Fred is in an experiential state E, then Fred knows or is in a position to know the essence of E.
(2) Fred is in experiential state E.
(3) The following is at least part of the essence of E: the experiential state E is identical to the physical state P.
(4) Fred does not know, and is not in a position to know, that experiential state E is identical to P.

These claims are inconsistent. If (1) and (2) are true, Fred knows the essence of E. But if (3) and (4) are true, Fred does not know the essence of E, that is, because to know the essence of E would be to know that it is identical to P, and Fred does not know that. On the other hand, (1) is an instance of the identification thesis; (3) follows from materialism as Lewis understands it; and (2) and (4) are the agreed on facts I mentioned. We can all agree, after all, that there are agents (or at least could perfectly well be) that both have experiences and are ignorant of their physical basis – Fred is just a stand in for such an agent. If (1–4) are inconsistent, one of them is false. But (2) and (4) are obvious, and (1) follows from what folk psychology says an experience is. Hence (3), and materialism generally, is false.

One might wonder whether (3) *does* follow from materialism as Lewis understands it. For example, one reason some philosophers find supervenience definitions of materialism (such as M3) attractive is that they apparently permit one to be materialist and deny an identity between physical states and psychological states. Be that as it may, for Lewis supervenience and identity naturally go together. Indeed, for him the idea that experience is defined in terms of its folk psychological role provides the starting point of a powerful argument for the identity theory; this is Lewis's famous argument for the identity theory set out in Lewis 1966 (see also Lewis 1972, 1994; and Armstrong 1968). The first premise of the argument is that, where "E" is some experiential state, E denotes the state that satisfies all or most the features above. The second premise is that, where "P" denotes some physical state, P is the state that satisfies all or most of the features above. The conclusion is that E is, that is, is identical to, P. The first premise here is supported by the approach to experience we reviewed earlier. The second premise is supported on empirical grounds – for example, on the ground that there is some state of the brain that is the state that satisfies most of the features above. And the conclusion of the argument follows, Lewis says (1972), "by the transitivity of identity."

The second argument against the cotenability of materialism and the existence of experience is the (better-known) *knowledge argument*. This argument may be set out in various ways, but a simple version has it as proceeding from two main premises. The first premise concerns what it is possible for a person to know; in particular, it is possible for a person to know all the physical facts as well as every fact that follows *a priori* from the physical facts, and yet not know what it is like to have an

experience of certain type. Jackson's (1982) Mary is the best-known illustration of this possibility, and the one that Lewis mostly focuses on.[7] The second premise of the knowledge argument is that if this is possible then materialism is false. The conclusion is that materialism is false, or anyway it is false if there are facts about what it is like to have certain experiences and if people know these facts.

One might wonder again whether this argument threatens materialism in the form that Lewis advances it. Another reason some philosophers find supervenience definitions of materialism (such as M3) attractive is that they apparently permit one to be materialist and deny the a priori deducibility of the mental by the physical. Such philosophers are *a posteriori* materialists rather than *a priori* materialists, as it is often put, and on the surface, the knowledge argument targets only the *a priori* materialist. Be that as it may, for Lewis supervenience and *a priori* deducibility naturally go together; he is an *a priori* materialist. The reason for this is his commitment to (and understanding of) the "two-dimensional" analysis of the necessary *a priori*. We will not go into the details of that position here. Suffice it to say that if materialism is true, and if M3 is the definition of materialism, there is a necessarily true conditional of the form "if S then S*," where S gives precise information about the instantiation of physical properties, as well as negative information about alien properties, and S* gives precise information about the instantiation of psychological properties. The issue that divides the *a priori* from the *a posteriori* materialist is whether this conditional is a priori or not. Lewis thinks that the "two-dimensional" analysis of the necessary *a posteriori* entails that the conditional "if S then S*" is *a priori* if materialism is true (1994, 297; see also Chalmers 1996 and Jackson 1998). Conclusion: *a posteriori* materialism is not a possible position.

33.5 Element #4: Distinct Conceptions of Experience

If he does not respond to the identification argument and the knowledge argument by rejecting identity or *a priori* deducibility, how does Lewis respond? In both cases, his strategy is, first, to distinguish two different conceptions of experience (or something related), and, second, to suggest that while the arguments rely on one conception, the materialist may employ another. The materialist need not insist that his own is the best one, or the perfect one, or the one that best answers to ordinary or philosophical thought; but he does need to argue that it a legitimate one.

This strategy is particularly clear in the case of the identification argument. As we saw above, if you operate with a conception of experience according to which something is an experience only if it satisfies all the requirements of folk psychology, it will follow that the identification argument is sound and materialism is false. But Lewis argues that a conception of experience that dispenses with the identification thesis is a legitimate one: something which satisfies the other features of experience listed above would still deserve the name "experience," even if one can imagine a more perfect deserver of the name. It is in this context that Lewis makes the remark about eliminativism that we quoted right at the beginning. Does Lewis deny the existence of experience? Yes and no; that is, he denies it if you mean something that satisfies all the requirements of folk psychology and the identification thesis in particular; he does not if you mean something that satisfies most of the requirements.

In the case of the knowledge argument, the "two legitimate conceptions" strategy assumes a slightly different form. Here Lewis focuses on two conceptions, not of experience, but of knowing what an experience is like. On the first, to know what it is like is to know a phenomenal fact, where a phenomenal fact is defined as a kind of fact which does not supervene on the physical facts, and which (remote cases aside) requires experience in order to be known – Lewis calls this the "hypothesis of phenomenal information." On the second, to know what it is like to see red is (merely) to know

how to do something, or to have a certain ability, namely, to recollect, recognize, and imagine certain experience – Lewis calls this "the ability hypothesis." Lewis first argues that the Mary example could be described in either way: on the first, she learns a phenomenal fact; on the second she learns how to do something. He then argues that if the example is described in the second way there is nothing in it to threaten materialism, since materialism does not rule out the possibility that someone might know all the physical facts, and yet gain some abilities.

Lewis's responses to these two arguments are different in an important way. In the case of the identification argument, both conceptions he operates with are present in folk psychology, or at any rate so Lewis claims. In the case of the knowledge argument, however, no such claim is made. In particular, the hypothesis of phenomenal information and the conception of knowing what it is like which goes along with it is an overt piece of philosophy – a "gratuitous metaphysical gloss" (1988, 290) – and indeed one which is question begging, since one can only know what an experience is like in this sense if materialism is false. This seems to me a weakness of Lewis's discussion of the knowledge argument in the 1988 paper. The psychological plausibility of the Mary example and the argument founded on it is very striking; it is implausible that only those who are in the grip of some tendentious metaphysics find them compelling. It would be preferable therefore if the two conceptions of "knowing what it is like" Lewis's response requires are plausibly part of ordinary thought.

In "Should a Materialist Believe in Qualia?" the 1995 paper that came out seven years after Lewis's main discussion of the Knowledge Argument, we find the materials to deal with this problem. In that paper, Lewis argues that "knowing what it is like" is a matter of knowing the essence of the experience; in other words, knowing what an experience is like is explicated in terms of the identification thesis. Moreover he argues explicitly that this idea is built into folk psychology, and is connected to the knowledge how that is involved in the ability hypothesis because according to folk psychology one has these abilities because one knows the essence. He also says that a commitment to the identification thesis is central to dualist thinking, and mentions Kripke (1980) in particular in this connection. While he does not explicitly draw the connection with the knowledge argument, it is natural to read his remarks as applying there too. If so, we may adjust his response to the knowledge argument in a way that avoids the problem about folk psychology just mentioned. According to this revised response, there are two conceptions of knowing what it is like: the first is merely to have the ability to imagine (etc.) the relevant experience, the second is to know the essence of that experience. The Mary example can be described in either way, but the materialist can explicitly adopt the first.

33.6 Know-How and the Ability Hypothesis

Turning now to objections to Lewis's account, the first I will consider focuses on Lewis's response to the knowledge argument. As we have seen, Lewis adopts the ability hypothesis here, and argues that Mary gains know-how or abilities rather than knowledge-that when she emerges from her cell. However, a number of philosophers (e.g. Loar 1990, Lycan 1996, Stanley and Williamson 2001, Stanley 2011; see also Alter 2001 and Cath 2009) have objected that knowing how to do something is simply a kind of knowledge that, and that in consequence the ability hypothesis is mistaken. The premise of this objection – that know-how is a species of know-that – is certainly plausible. And no one can deny that Lewis formulates the ability hypothesis in terms of know-how which is not know-that. The question is whether this part of his account might be dropped while key elements of the ability hypothesis are retained – whether there is a "fallback," as Stanley and Williamson (2001) put it.

The problem is that there is an apparently decisive argument that no fallback is possible. For suppose that knowing how to imagine (etc.) an experience is a kind of knowledge-that – for example,

knowledge of some proposition *p*. Either Mary knows *p* in her room or she does not. Suppose she does. In that case, a proponent of the ability hypothesis may still say that Mary gains some ability on coming out, for it is possible to separate the ability from the knowledge-how.[8] But what the proponent can't say is that in gaining this ability Mary gains some knowledge-how – for by assumption she already had that knowledge. And this means that the ability hypothesis loses a lot of its appeal. Central to the Mary case is the idea that Mary learns something on coming out, (i.e. gains some knowledge on coming out). If the ability hypothesis can explain this by saying Mary gains some knowledge how, well and good; but if the ability hypothesis cannot explain it, it provides no reasonable description of the case. Suppose then that she does not know *p* in her room. In that case, a proponent of the ability hypothesis can say that Mary gains some propositional knowledge on coming out, and may appeal to this to explain the knowledge intuition. But the problem now is that the propositional knowledge in question cannot be (if materialism is true) the sort of propositional knowledge that Lewis is mainly interested in, that is, the sort that distinguishes one possible world from another. It is true that one can define various notions of propositional knowledge that do not distinguish one possible world from another. But Lewis is scathing on the idea that doing so is connected to the Mary example. Indeed of the six ways he sets out (in Lewis 1988) of missing the point of the knowledge argument, no fewer than four are related to this idea!

Yuri Cath (2009) has suggested nevertheless that the second option here is the best one. His idea is that pre-release Mary might have the propositional knowledge but not under the right mode of presentation. Cath makes a good case that this is plausible for some proponents of the ability hypothesis, but it is hard to see it as plausible for Lewis, as indeed Cath points out at the end of his paper. If we are out to find a plausible fallback for Lewis, it is better to focus on the first horn of the dilemma just outlined, and to try to defend the view that the ability Mary gains is genuinely a kind of knowledge even if it is not know-how.

I think such a defence emerges when we observe that a sentence such as "Mary knows what it is like to see red" is ambiguous in a way that many "know what" sentences are. On the one hand, it might be used to say that Mary knows some fact that (in the context) answers the embedded question "What is it like to see red?" Since that question seems intuitively to ask "What type of experience does one have when one sees a red thing?" the sentence on this reading means that Mary knows some fact that (in the context) answers this question. Call this the "interrogative" reading of the sentence. On the other hand the sentence can also be used to say that Mary knows the thing or property denoted by the referring expression "what it is like to see red." Since the referring expression intuitively denotes a type of experience, on this reading the sentence says that Mary knows a type of experience, that is, the one you have when you see red. Call this the "free relative" reading of the sentence. It is the interrogative reading we use when we say of pre-release Mary "she wonders what it is like to see red" – what she wonders is not a type of experience, but what fact answers a certain question. It is the free relative reading we use when we say of post-release Mary, "she loves what it is like to see red" – what she loves is a type of experience, not an answer to a certain question.[9]

Which of these readings is Lewis interested in? While he does not explicitly draw this distinction, I think it is plausible to interpret him as being concerned mainly with the free relative reading. One data point is that Lewis explicitly says (1988, 286, fn.12) that the phrase "what experience E is like" in the larger phrase "know what experience E is like" denotes E itself; this strongly suggests he is setting aside the interrogative reading of the sentence. Another data point is Lewis's second way to miss the point of the knowledge argument (1988, 266). The second way to miss the point, according to Lewis, is to suppose that what Mary doesn't know is an answer to a certain question, namely, "what type of experiences she will have when sees red for the first time." Lewis argues that she does know an answer to this question, and so "knowing what it is like" should not be understood in this interrogative way.

Suppose then that Lewis does indeed employ "know what it is like to see red" on its free relative reading; in that case, we have a response to the second horn of the dilemma outlined earlier. Mary may well know how to imagine red but she does not have the ability to do so until she has an experience. But why is the ability that she gains properly called a state of knowledge? The answer is that it is the state of knowing a type of experience, that is, "knowing what it is like" on the free relative sense.

33.7 Contextualism and the Identification Thesis

The second objection I want to consider focuses on the identification hypothesis. We have seen that Lewis rejects this, but a number of philosophers (Schwarz 2007; Stalnaker 2008) have pointed out that doing so seems inconsistent with Lewis's own contextualist account of propositional knowledge.

According to this account, "subject S knows proposition P iff P holds in every possibility left uneliminated by S's evidence; equivalently, iff S's evidence eliminates every possibility in which non-P" (Lewis 1996, 422). This account raises a number of questions, but the connection to the identification thesis has to do with Lewis's account of what it is for evidence to eliminate every possibility.[10] Here is what he says:

> A possibility W is uneliminated iff the subject's perceptual experience and memory in W exactly match his perceptual experience and memory in actuality. (1996, 424)

Pairing this account of the elimination of possibilities with the main part of Lewis's contextualism already has an important consequence for Lewis's account of experience. In particular, it entails the self-intimation thesis about experience mentioned earlier: that if I am in a certain experiential state, I know that I am in that state. Some might think that this is bad enough,[11] but Stalnaker and Schwarz argue that the problem goes deeper. As Stalnaker puts it:

> . . . on Lewis's account, we will at least know, in any context, that the possibilities excluded by our experience – possible situations in which our experience does not match our actual experience – are possibilities that are incompatible with our knowledge. The problem is that all of these possibilities will be possibilities in which our experience has whatever essential properties our actual experience has. That is, Lewis's account of knowledge implies that even in our most sceptical context, we will know the essential nature of our experiences. . . . Lewis's account of knowledge entails the identification thesis that he rejects . . . (2008, 100)

To bring this out, suppose I am in an experiential state E; and suppose in accordance with materialism of the kind Lewis defends, that this state is identical to some physical state P. As we have just seen, it will immediately follow that I will know that I am in E. Does it follow also that I know that I am in P? According to the objection, the answer is yes. For in all the worlds that remain uneliminated by my evidence I will be in P; hence I will know that I am in P. And this seems tantamount to the identification thesis.[12]

It is hard to overstate what a disaster it would be for Lewis's account of materialism and experience if this objection were correct. As we have seen, the rejection of the identification thesis is crucial to his response to the identification argument, and perhaps too the knowledge argument. But is the objection correct? I think this depends on what it is for a subject's perceptual experience in one world to "exactly match" that subject's perceptual experience in another. The first thing to say is that what Lewis has in mind here are events of experiencing, and as we saw an event is an experience if it

consists in the onset or persistence of an experiential state. Presumably, therefore, experience e in possible world W will match experience e* in possible world W* if and only if they consist in the onset or persistence of the same experiential state; for short, if and only if they consist in the same experiential state. But what it is for two experiences (in distinct possible worlds) to consist of the same experiential state? I think there are two possibilities here. According to the first, two experiences consist of the same experiential state (and so match each other) if and only if they consist of experiential states with the same functional role. According to the second, two experiences consist of the same experiential state (and so match each other) if and only if they consist of experiential states with the same essence. Two experiences could match each other in the first sense and not in the second. For example, contrast a dualist world D and a materialist world M, and suppose that at both worlds I am in pain. Do the two experiences I have here (i.e. the one I have in D and the one I have in M) match? According to the first account of matching, they do; but according to the second they do not.

The distinction between these two accounts of matching provides the answer to the objection made by Stalnaker and Schwarz.[13] It is true that if the second account of matching is in play, it will follow that I know that I am in P, and this is objectionable. But if only the first account is in play, nothing similar is true: in some of the possible worlds at which I have an experience which matches the actual experience I will not be in P. Is it reasonable to think that Lewis is operating only with the first account? It would seem so. First, as we saw earlier, an experiential state is something defined in terms of its folk psychological role, which suggests that it is not defined in terms of its essence. Second, the contextualist program in epistemology he outlines could proceed unhindered if only the first account is in play. If so, the objection fails and disaster is avoided.

Concluding Remarks

I have defended Lewis against these two objections, but I should say for the record that I have a number of sharp disagreements with Lewis's account of experience and materialism, which I have set out elsewhere. I will conclude with a brief statement of what these are.

Disagreement #1 concerns the ability hypothesis. Lewis's ability hypothesis focuses on the idea of a novel experience – that is, an experience one has not had before – and this focus is shared by many contemporary accounts of the knowledge argument, in particular acquaintance views, phenomenal concept views and so on. Now the Mary example *is* an example in which someone has a novel experience. But this feature of the example is incidental to the persuasiveness of the argument founded on it; other examples can be constructed that support that key possibility just as well but which do not involve novel experiences. The ability hypothesis has nothing to say to examples of this kind, and neither does any response to the knowledge argument founded on novel experiences (see Stoljar 2005, 2006, chapter 10). So I think a completely different approach is required.

Disagreement #2 concerns the identification thesis. I agree with Lewis that the identification thesis is false, and in that sense with his response to the identification argument. But I doubt that the identification thesis plays the role in ordinary and philosophical thought that Lewis thinks it does. For one thing, what is built into folk psychology is an empirical question, logically like the question of whether folk physics is a version of medieval impetus physics. But even so, the suggestion that the identification thesis is built into folk psychology is something about which we should be sceptical. It is plausible that folk psychology asserts some connection between experience and belief about or knowledge of experience. But there are many such theses, and the identification thesis is only one (and a fairly extreme one at that); there is so far as I know no good reason to believe that this thesis *in particular* is built into folk psychology (see Stoljar 2009). Nor is it credible to suppose that the

identification thesis is a presupposition of philosophical thought that is sympathetic to dualism. Take Kripke's discussion of the modal argument in *Naming and Necessity* (1980), something that Lewis mentions. Kripke is concerned there to emphasize that there is no distinction between feeling pain and being in pain, but I think one can emphasize that without being committed to the identification thesis (Stoljar 2006, chapter 11, 2009).

Disagreement #3 concerns Lewis's optimism about current science. We saw earlier that, according to Lewis, a property is a physical property if and only if it is expressed by a physical theory that is a "presumably somewhat improved" version of what we currently have. I think, as I said, that this is subject to persuasive counterexamples (see Stoljar 2010, chapter 4). But there is a more important issue at stake here. For Lewis's discussion of physical properties and indeed of materialism itself embodies a quite radical optimism about current science, an optimism that is routinely underappreciated, as Jackson noted in his first discussion of the knowledge argument (see Jackson 1983).[14] Lewis himself thought that physics was very nearly complete, that is, that every type of physical truth was known or nearly so, and that every contingent truth follows *a priori* from physical truths. Now the question of whether optimism or pessimism is true is an example of a disagreement in philosophy that will not easily be resolvable, if at all. Presumably, the dedicated optimist will remain unmoved by reminders about the fallibility of humans, and about the history of wrong turns and mistakes in our attempts at understanding the world (see Stoljar 2010, chapter 5). Speaking personally, however, I find this sort of optimism intellectually alien. I don't think scientists believe it, even if they occasionally say things that suggest that they do. And, if they do believe it, I think they are wrong; I don't think it is worthy of belief. Moreover, even if this sort of optimism is by some miracle true, I don't think arguments that presuppose it are persuasive.

Disagreement #4 concerns what the knowledge argument and similar arguments do presuppose. As I have said, in my view the knowledge argument does not presuppose the identification thesis, nor is it essentially focused on novel experiences. However, I do think it presupposes something like the optimism just mentioned, in particular, I think it presupposes that we have complete knowledge of the physical or non-experiential world, if not in detail then at least in outline (Stoljar 2006). But as I have just indicated, that presupposition is in my view false (in fact I think the plausibility of these arguments is evidence that it is false). And if it is false, I think it is quite clear where the arguments go wrong. If we assume we have incomplete rather than complete knowledge of the relevant facts, what the knowledge argument and related arguments show is the unremarkable truth that experience comes apart from some physical facts (i.e. the known ones), not the remarkable truth that it comes apart from all.

I know these points are controversial; my aim in this last section was to state my disagreements with Lewis, not defend them.

Acknowledgments

I am grateful for comments from Jed Buchanan, Daniel Nolan, Paulo Santorio, and Wolfgang Schwarz.

Notes

1 Some might prefer "physicalism" to "materialism," but Lewis did not and I will follow him here. For a discussion of the issues lying behind the phraseology, see Stoljar 2010, chapter 1.

2 As Paolo Santorio pointed out to me, there is at least the following problem with this definition: it apparently counts as alien to a world W any property which meets (1) and (2) and yet could be instantiated in worlds that have the same laws as W. I will set this issue aside here.

3 Lewis in fact discusses the view that all natural properties are physical properties, rather than all fundamental properties, but I will take him as talking about the fundamental properties definition too – the reference to "quarks" in the passage at least is suggestive of it.
4 Examples like this may be constructed by considering "twin-physics" cases. See Stoljar 2010, chapter 4 and the references therein.
5 Some of these principles are more naturally expressed in terms of experiencings, and some are more naturally expressed in terms of experiential states. I will leave that unresolved in the text.
6 In (1972) Lewis says that self-intimation is an open question, whereas in (1996) he is committed to it, as we will see below.
7 Here is Lewis's own description of the case:

> Mary, a brilliant scientist, has lived from birth in a cell where everything is black and white. (Even she herself is painted all over.) She views the world on black-and-white television. By television she reads books, she joins in discussion, she watches the results of experiments done under her direction. In this way she becomes the worlds leading expert on color and color vision and the brain states produced by exposure to colors. But she doesn't know what it is like to see color. And she never will, unless she escapes from her cell. (1988, 263)

8 It is possible for someone to know how to do something without having the ability to do it; witness the unfortunate pianist who knows how to play but lacks the ability because she has just lost her arms in a car crash (cf. Ginet 1975; Chomsky 1988; Stanley and Williamson 2001; Cath 2009) Similarly, it is possible that Mary knows to imagine but does not have the ability, that is, because she has not had the experience.
9 For the distinction (plus the labels), though not the application to "know what it is like," I am indebted to Schaffer 2010.
10 As Wolfgang Schwarz pointed out to me, it is Lewis's account of what it is to eliminate possibilities that generates this problem, not the contextualism, strictly speaking. I will ignore this is in the text however.
11 As Hawthorne (2004, 60, n.26) points out, "if I have a visual array with 137 red dots, it does not seem that I automatically know that," though Lewis's account apparently implies that he does.
12 Seems tantamount, but perhaps is not quite identical; the identification thesis requires not simply that I know that I am P but that I know that E is P. However, the knowledge described in the text is bad enough.
13 In his (2007), Schwarz suggests something along these lines, but objects that, on that view, one will always know various causal facts involved in having an experience. I am not sure if that is an objection or not, but even if it is it is clearly preferable to knowing the essence of the experience from Lewis's point of view.
14 This optimism is distinct from Lewis's (2009) argument for Ramseyan Humility, as Lewis notes in footnote 5 of that paper.

References

Armstrong, D.M. (1968). *A materialist theory of the mind*. London: Routledge.
Alter, T. (2001). Know-how, ability and the ability hypothesis. *Theoria*, 67(3), 229–39.
Cath, Y. (2009). The ability hypothesis and the new knowledge how. *Noûs*, 43(1), 137–56.
Chalmers, D. (1996). *The conscious mind: in search of a fundamental theory*. New York: Oxford University Press.
Chomsky, N. (1988). *Language and problems of knowledge*. Cambridge, MA: MIT Press.
Ginet, C. (1975). *Knowledge, perception and memory*. Dordrecht: Reidel.
Hawthorne, J. (2004). *Knowledge and lotteries*. Oxford: Oxford University Press.
Jackson, F. (1982). Epiphenomenal qualia. *Philosophical Quarterly*, 32, 127–36.
Jackson, F. (1998). *From metaphysics to ethics: a defence of conceptual analysis*. Oxford: Clarendon Press.
Kripke, S. (1980). *Naming and necessity*. Cambridge, MA: Harvard University Press.
Lewis, D. (1966). An argument for the identity theory. *Journal of Philosophy*, 63, 17–25. Reprinted in Lewis (1983b), pp. 99–107. All references are to the latter version.

Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, 5, 249–58. Reprinted in Lewis (1999), pp. 248–61. All references are to the latter version.
Lewis, D. (1978). Mad pain and Martian pain. In N. Block (Ed.), *Readings in the philosophy of psychology*, Vol. 1, pp. 216–22. Cambridge, MA: Harvard University Press. Reprinted in Lewis (1999), pp. 325–31. All references are to the latter version.
Lewis, D. (1983a). New work for a theory of universals. *Australasian Journal of Philosophy*, 61, 343–77. Reprinted in Lewis (1999), pp. 8–55. All references are to the latter version.
Lewis, D. (1983b). *Philosophical papers*, Vol. I. Oxford: Oxford University Press.
Lewis, D. (1983c). Individuation by acquaintance and by stipulation. *Philosophical Review*, 92 (1), 3–32. Reprinted in Lewis (1999), pp. 291–324.
Lewis, D. (1988). What experience teaches. *Proceedings of the Russellian Society*, 13, 29–57. Reprinted in Lewis (1999), pp. 262–90. All references are to the latter version.
Lewis, D. (1994). Reduction of mind. In S. Guttenplan (Ed.), *A companion to the philosophy of mind*, pp. 412–31. Oxford: Blackwell. Reprinted in Lewis (1999), pp. 291–324. All references are to the latter version.
Lewis, D. (1995). Should a materialist believe in qualia? *Australasian Journal of Philosophy*, 73, 140–44. Reprinted in Lewis (1999), pp. 325–31. All references are to the latter version.
Lewis, D. (1996). Elusive knowledge. *Australasian Journal of Philosophy*, 74, 549–67. Reprinted in Lewis (1999), pp. 418–46. All references are to the latter version.
Lewis, D. (1997). Naming the colours. *Australasian Journal of Philosophy*, 75, 325–42. Reprinted in Lewis (1999), pp. 332–58. All references are to the latter version.
Lewis, D. (1999). *Papers in metaphysics and epistemology*. Cambridge: Cambridge University Press.
Lewis, D. (2009) Ramseyan humility. In Nola & Braddon-Mitchell (2009), pp. 203–22.
Loar, B. (1997). Phenomenal states. In N. Block, O. Flanagan & G. Guzeldere (Eds.), *The nature of consciousness*, pp. 597–616. Cambridge, MA: MIT Press.
Lycan, W.G. (1996). *Consciousness and experience*. Cambridge, MA: MIT Press.
Nola, R. & Braddon-Mitchell, D. (Eds.) (2009) *Conceptual analysis and philosophical naturalism*. Cambridge, MA: MIT Press.
Schaffer, J. (2010). Knowing the answer redux. *Philosophy and Phenomenological Research*, 78(2), 477–500.
Schwarz, W. (2007). Lewis on knowing one's evidence. Blogpost, November 6, http://www.umsu.de/wo/2007/500, accessed June 8, 2014.
Stanley, J. (2011). *Know how*. Oxford: Oxford University Press.
Stanley, J. & Williamson T. (2001). Knowing how. *Journal of Philosophy*, 98, 411–44.
Stalnaker, R. (2008). *Our knowledge of the internal world*. Oxford: Oxford University Press.
Stoljar, D. (2005). Physicalism and phenomenal concepts. *Mind and Language*, 20(5), 469–94.
Stoljar, D. (2006). *Ignorance and imagination*. Oxford: Oxford University Press.
Stoljar, D. (2009). The argument from revelation. In Nola & Braddon-Mitchell (2009), pp. 113–38.
Stoljar, D. (2010) *Physicalism*. London: Routledge.

Part VI

Ethics and Politics

34

Lewis on Value and Valuing

PETER RAILTON

34.1 Introduction

Sometime in the late 1980s, David Lewis was on a visit to Ann Arbor and I was escorting him to a well-known hobby shop in town. Those were the days – as things turned out, the sunset days – of hobbies as a highly visible part of mainstream culture. Before we all got "busy."

Like many of the most creative people I have known, Lewis never seemed "busy." He always seemed to have time to stop and noodle something over, to drink and sing on into the night, to launch off on a walk in search of some curiosity, or simply to stand quietly at a gathering, contented and bright-eyed, thinking.

That afternoon, Lewis wasn't busy, so we took the long walk to the hobby shop and back, in pursuit of a bit of miniature rolling stock for his model train set. Along the way we started talking about future directions in our work. He wasn't yet sure where he'd be going next, but he was sure it would be something quite different. "Where? Let's see now . . . ," he answered, rubbing his wispy hair. He took his time, and seemed, as he often did at such moments, to be looking up a complex expected utility function. The first result: "Here's one thing I'm sure of. Not conditionals. Not counterfactuals. Not causation. Nothing like that." Then: "Value," he said empathically, with a short laugh and an unmistakable, slightly mischievous, twinkle in his eyes. It was clear that he was delighted to have given himself the scope for this.

In a way, Lewis was ideally equipped for the venture. In his life he was a great celebrator of value – in ideas, arguments, music, history, trains, and, above all, sociability and humor. So it was natural that when his theory of value emerged, it would be a pluralistic, open-ended theory that placed desire and enjoyment, not judgment and norms, at the core. Indeed, I suspect that, in his own life, desiring and valuing, and valuing and desiring, were intimately connected.

Lewis was fond of saying that he would have preferred to be a piecemeal philosopher rather than a system-builder, but that he kept finding that the ideas he had developed in one area helped him to see his way in another. A basic perspective and project emerged unbidden, "Humean supervenience": the idea that worlds are just mosaics of particular facts – "one darn thing after another" – and that

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

whatever we accept by way of causal relation, natural law, explanation, probability, possibility, or necessity supervenes upon these mosaics. Fix these facts, and you will have fixed the rest. "Rubbish" such as causal or probabilistic propensities, necessities as metaphysical glue, or ghostly sense-data, was not only epistemically dubious, but irrelevant even if it existed – it could cause nothing (given the nature of causation), and explain nothing (given the nature of explanation). It was otiose in every sense.

Instead of objective chance in the form of probabilistic propensities, for example, Lewis constructed "objectified subjective probability" (Lewis 1980). Unlike chancy propensities as *things-in-themselves*, objectified subjective probabilities would have the *a priori* relation to the axioms of probability and the norms of rational belief formation and choice that would fit them to enter into a thinking person's account of the world and the conduct of her life.

Others, myself included, were less happy with this approach. Though the interpretation of quantum mechanics and quantum field theory remains problematic, it does seem that these theories as standardly understood make a strong case for lawful chance as a property of physical systems – chance "in the world" – distinct from patterns of individual events, yet responsible for them. This makes at least conceivable the idea of objective probabilistic propensities that involve no detour through rational credences or their *a priori* constraints, and opens up the problem for Humean supervenience of worlds with the same sequence of particular facts, but different lawful chances. (For a reply to such worries, see Lewis 1986.)

However that may be, many philosophers unconvinced by Humean supervenience for matters of law, causation, or chance find such a view highly plausible in the case of value. It seems perfectly natural to think that value is introduced into the world by the appearance of subjectivity – of beings to whom something matters. The challenge is to understand this notion of mattering, and how it relates to value, but, to philosophers of diverse stripes, subjective attitudes seem like a good place to start.

34.2 From Valuing to Value

Thus, Lewis begins his contribution to "Dispositional Theories of Value"[1]:

> **(1)** Roughly, values are what we are disposed to value. (114)

For familiar reasons, this rough gloss is too rough. As a shy teenager, my fear of change might dispose me to value remaining forever in the sheltered environment of my family and familiar community, and to disvalue the idea of ever leaving this microcosm to try the wider world. However, if I lack the experience to see that living in the wider world would be quite different from how I imagine it, and, indeed, if a more vivid and accurate imaginative acquaintance with what such a life would actually be like would induce in me a strongly favorable attitude toward it, this seems like a good reason for me to question my present disvaluing of it.

So Lewis qualifies (1). "Less roughly," he writes, "we have this schematic definition":

> **(2)** *Something of the appropriate category is a value if and only if we would be disposed, under ideal conditions, to value it*. (114, emphasis original)

Here, then, we have a form of "objectified subjectivity" – a subjective response under ideal conditions.

Theories of this general sort are sometimes called "qualified attitude" accounts, and the history of such accounts of value stretches back at least as far as Hume on aesthetic value and Mill on utility (Hume 1985; Mill 1962). But Lewis's position is more democratic than either Hume's or Mill's, since the attitudes upon which he fixes are not the attitudes of someone with greatest expertise, full knowledge, widest experience, highest sensitivity, and least prejudice. Rather, they are the attitudes of the particular, partial individual in question – informed in a rather specific way, namely, by fullest and most vivid possible imaginative acquaintance with the very possibility whose value is under consideration. "Complete non-evaluative information," "ideal sensitivity," or "impartiality" are not relevant, Lewis thinks, for ideal *valuing* – these notions come into play only when we need to think about ideal *judging*, *weighing*, or *balancing* of values (126). If, as a result, values turn out to be a more motley, conflicting, lowbrow, and narrow-minded lot than philosophers have wanted to think – so be it. Let's not deny a value its day in the sun just because we don't like its looks, or because we can see more weighty values that overshadow it.

Let's spell out a few of the more salient features of Lewis's account.

It is reductionist, in at least three ways. First, unlike many qualified attitude theories, the effect of fullest possible imaginative acquaintance upon dispositions to value that interests Lewis is a strictly *causal* effect – value is a matter of those dispositions to value *induced* by imaginative acquaintance, not those dispositions or attitudes such an acquaintance would *warrant* or *merit*. In this respect, his account is closer to traditional secondary-quality theories than most contemporary qualified attitude theories (see D'Arms & Jacobson 2000). It promises to take us out of the circle of normative concepts.

Second, this promise is strengthened by the way that Lewis then distances himself from traditional secondary-quality theories. Consider first a rough secondary-quality definition of *red*:

(3) *X* is red $=_{\text{def}}$ *X* is disposed to produce a red sensation in normal human observers when attended to visually under standard conditions.

Lewis begins with:

(4) *X* is a value $=_{\text{def}}$ we are disposed to value *X* under ideal conditions.

(4), like (3), replicates the term under analysis in the analysandum. Most have thought this irreducible – there is no other way of characterizing the distinctive experience red vs. blue or green objects are disposed to produce. Lewis rejects accounts of the valuing attitude in terms of *judging to be valuable*, *taking to be valuable*, *believing to be valuable*, or even *experiencing as valuable* (114). Near enough, he believes, can be defined in terms of what we *desire to desire* – at least when these desires are understood as non-instrumental. So (4) quickly becomes:

(5) *X* is a value $=_{\text{def}}$ we are disposed to desire to desire *X* intrinsically under ideal conditions.

Third, although we might suspect – especially in light of the history of qualified attitude theories – that (5) continues to harbor normative notions in the idea of "ideal conditions," Lewis, we noted, believes it can be fully spelled out without them:

(6) *X* is a value $=_{\text{def}}$ we are disposed to desire to desire *X* intrinsically under conditions of fullest possible imaginative acquaintance with *X*.

"Fullest possible" is just that – what humans can accomplish by way of vivid and accurate simulation of what it would be like for *X* to obtain. Lewis writes of his account, "It reduces facts about value to facts about our psychology" (113).

This can sound odd. It locates values in us – more precisely, in our dispositions under ideal conditions – not in the objects of our evaluative attitudes. One might see this as a "classically" Humean view. We can speak as if our valuings were responses to things in the world:

> Can we say that the valuing ensued because *X* was a value? – Maybe so, but if we do, we are not saying much: it ensues because there is something about imaginative acquaintance with *X* that causes valuing. (122)

But this is really a fact about the causal effects of *our imagining X*, not about *X* as such: "values themselves aren't disposed to do anything" (124, n.). So valuing isn't about organizing perceptual experience under normative concepts, or about detecting non-natural properties "in the world", rather, it is about our capacity to be engaged in favor of certain possibilities. Learning about value is learning about these capacities. That can be hard: *noce te ipsum* is no easy task. But we have a "canonical" way of gaining knowledge of our dispositions to value, namely, by vividly simulating what it would be like to realize some putative value. And we have various aids, such as good novels and the advice of others. To be sure, Lewis notes, anyone's actual imagining can fall short of fullest possible imaginative acquaintance in many ways, but if, improbably, someone could "gain the fullest imaginative acquaintance that is humanly possible, then, I suggest, his valuing would be an infallible indicator" (121).[2]

Of course, talk of "infallible indicator" rather suggests that dispositions to value are not themselves the stuff of value, but rather the way we home in on them. Lewis, however, wants the stronger view. He recognizes that this places his account in uncomfortable company – behaviorism about mental states and phenomenalism about the external world both equate what we normally take to be indicators with the things in themselves (131). For the case of value, however, Lewis thinks there is no wedge we could use to pry the *indicator* from the (putative) *indicatee*, the way that we can readily and in convincing "corroborative detail" imagine behavior that cleverly counterfeits mental states or a world that produces systematically inaccurate sense experience (132). Something *not* of value that elicits in us intrinsic desire to desire when we have the fullest and most vivid possible imaginative acquaintance with it? This sounds to Lewis like something that is *not really moving* even though, when we have the fullest possible imaginative acquaintance with it, it moves profoundly us in its own right.

Lewis's idea that "imaginative acquaintance is all we need" (123) might still raise some questions. A distinctive element of his theory is that, unlike most ideal attitude theorists, he does not give a *generic* characterization of the ideal conditions for our responses. Such a generic characterization is usually something like "full non-evaluative information," "wide experience," "disinterested," etc., which has the advantage that it is not tailored to any specific putative value. Lewis takes the opposite tack:

> . . . when I say that *X* is a value iff we are disposed to value *X* under ideal conditions, I do not mean conditions that are ideal *simpliciter*, but rather conditions that are ideal *for X*. (126)

It might seem implausible to say that *X* is a value when fuller knowledge of *X*, beyond an ideally vivid simulation of what it would be like, would undermine desire. Suppose, for example, one learns *how* the simulation of *X* in particular gives rise to one's desire to desire it. Perhaps the fact that simulating *X* engages one motivationally in this way depends upon a path of causation that is routed through implicit prejudices, phobic responses, behavioral conditioning, or some neurological glitch. Lewis writes:

> If some feature of our lives originated by kin selection, or Pavlovian conditioning, or sublimation of infantile sexuality, that is irrelevant to what it is like in itself. Unless he can overcome the illusion of relevance, a valuer will be more reliable if he remains ignorant of such matters. (125)

We can certainly agree with Lewis that the "genetic fallacy" must be resisted. Our notion of beauty in living form, some evolutionary psychologists theorize, has to do with the information that bilateral physical symmetry carries about the current health and past history of vulnerability to disease of potential mating partners. So be it. But only the most fanatical Platonist would, on this ground, become a nihilist about beauty.

However, Lewis makes a stronger claim:

> How does imaginative acquaintance cause valuing, when it does? How does imagination render values attractive? Does it happen the same way for all values? – For our purposes, it is enough to say that it happens. We needn't know how. (122, n.)

It seems to me that it would difficult to sustain the thought that we don't need to know anything about how sensory experience is typically caused, or how it in turn typically causes belief, in order to do normative epistemology, or to worry about whether knowledge is present in the case of atypical or "devious" causation. Should matters be different in the case of value?

Here is what strikes me as an easy case in which causal history matters, even for someone with an overall picture of value like Lewis's. Let's say that, under fullest possible imaginative acquaintance with *X*, we desire to desire it. However, we later learn from theoretical psychological investigation that, in the particular case of *X*, fullest possible imaginative acquaintance with *X* cannot take us very close at all to what it would *actually* be like to live an *X*-ish life. Moreover, suppose that actually living an *X*-ish life would lead to disvaluing it in its own right. Is *X* genuinely a value, or is it only an apparent value? The intuitive sense inclining me toward denying *X* the status of value here is no different, I think, from the intuition with which we began – of a young Peter mistaken about value because he cannot imagine what it would be like to live in the larger world. It seems to me that fuller information can be relevant not only for purposes of weighing value – it might be necessary to dispel illusions about how we come to have certain values, or inform us of the limitations of imagining's power.

34.3 Pluralism about Value

Imagining can have limits in other ways as well, as Lewis recognizes. No one of us, perhaps, is capable of imagining all possible values (126). This might arise if the psychic development involved in coming to imagine fully and vividly what it is like to be an honor-bound Achilles-like warrior glorying in bloody battle preclude those involved in similarly imagining what it is like to live the self-denying, pacific life of a Quaker, and vice versa. "It might happen that imaginative acquaintance with *X* would leave traces, in one's valuing or otherwise, that got in the way of afterward imagining *Y*" (127). Looking at the longer sweep of history and culture, which values will ever be within the scope of actual humans is bound to be contingent upon the particular trajectories of their individual or shared lives, and the processes involved might be irreversible.

Lewis's response to this problem tells a lot about him:

> . . . it might happen that there were values that could not even be valued at once. If so, then conflict of value would go deeper than is ever seen in hard choices; because what makes a choice hard is that conflicting values *are* valued together by some unfortunate chooser. An alarming prospect! – or exhilarating, to those of us who delight in the rich variety of life. (126)

Lewis makes no assumption that all humans will have the same response to imagining any given putative value *X* as fully as possible. We humans are a lot alike, but we are also each unique in various ways. Lewis is thus "wary of presupposing that all mankind are alike in their dispositions to value" (127). Life experiences can shape us profoundly, opening us to some values but closing us to others. Most of us, I suspect, do not think that Quaker values are spurious because we simply cannot interest Achilles or Nietzsche in them – even under conditions of ideal imaginative acquaintance. Moreover – if we eschew Pollyanna-esque wishful thinking – our own inability, given our particular trajectories, to enter into the spirit of humiliating the weak or massacring our enemies as ends in themselves should not lead us to pronounce these as beyond the pale of human value.

Of course, a philosophical theory might be able to avoid some such inconvenient values by insisting that the only genuine values are those that, necessarily, all rational beings would converge upon. But it would avoid embarrassment only by running a nearly certain risk of denying the reality of value altogether – Quaker or warrior. Such an account is, for Lewis, simply an invitation to an error theory, and "The fire is worse than the frying pan" (134).

Such a judgment is a trade-off, a Lewisian philosophical "cost-benefit" calculation that is prepared to give up "requiring values to be all that we might wish them to be" in order to have an account that reflects what seems to matter most in our discourse and practice involving value. For example, philosophers known as "judgment internalists" have argued that there is a conceptual connection between making a sincere value judgment and being motivated to act accordingly (Darwall 1997). It would betray a failure to grasp the concept of value, on such a view, to say that warrior societies grounded upon honor have a value that leaves one cold.

On a Lewisian view, accepting such possibilities could simply be a recognition that, as things stand, one lacks the capacity to fully appreciate what a warrior's life is like – what might be seen in it – along with a certain humility about projecting one's current valuings to all places and times.

> Our theory makes a place for truth, and in principle for certain knowledge, and in practice for less-than-certain knowledge, about value. But also it makes a place for ignorance and error, for hesitant opinion and modesty, for trying to learn more and hoping to succeed. (123)

Lewis locates the flaw in "judgment internalist" theories exactly here – by tying evaluation essentially to one's current motivations, they make these responses dispositive:

> One fault of some subjective and prescriptivist theories is that they leave no room for modesty: just decide where you stand, then you may judge of value with the utmost confidence! (123)[3]

There *is* a connection between valuing and being motivated, for Lewis, but it is iffy, and goes via imaginative acquaintance and the "canonical method" of learning about value, rather than acts of judgment. This is a kind of "existence internalism" – for *X* to be a value, we must, under suitable conditions, be disposed to desire to desire *X*. Plenty of room for error, uncertainty, and incompleteness here. Is our imaginative simulation as full as possible? If it were fuller, would it cause us to have different desires? Have we attended to the simulation carefully enough in its own right, or might distraction, fatigue, or competing desires be keeping us from fullest possible engagement with it? And, of course, who are "we"? Lewis writes: "Nothing less iffy would be credible. But it is still less credible that there is no connection at all" (116). Here, too, Lewis is striking a balance. It might be pleasant to dream of satisfying both existence and judgment internalist intuitions, but when a full elaboration of theory shows how they come apart, we have to ask which, if either, is most fundamental to our thought and practice, and take sides.

Lewis wants value to have a credible epistemology – not to come for free along with one's convictions, whatever they might be. He wants learning what is valuable to be suitably disciplined *a posteriori* by relevant facts and procedures, but this typically assigns little or no weight to how a person, or his ideology or religion, currently conceives value. Lewis emphasizes that it is real work for someone to know what has value (127), and that this corresponds to how we actually think about things. The School of Hard Knocks and the process of sharing our thoughts with others are better at revealing what is valuable in life than *a priori* intuition or existential commitment.

34.4 Some Questions

I find Lewis's account of value congenial in many respects, and I hope my discussion thus far has helped make it plausible. It is an important contribution to the literature on value, though, like all such contributions, it raises a host of further questions. Lewis himself saw it as the beginning of an account, not the last word. His untimely death denied us its further development. I was fortunate to have had the chance to discuss some questions of further development with him, and the remarks that follow reflect in part those conversations. I will focus on three areas.

34.4.1 Valuing

It seems fair to say that that which I desire to desire is something like a *settled preference* of mine. But that seems different from *valuing* – even when the preference is non-instrumental.

Ordinarily, we distinguish between what we like or like to like, on the one hand – say, a fanciful but contented liking for the commercial jingles of the 1950s, or a settled preference for doing the bills on Sunday morning – and what we value or claim to have value, on the other. "I know what I like, and I like *that*" is one way of resigning without a fight a discussion of the aesthetic merits of one's peculiar way of furnishing one's room. "I know what I like, and I *don't* like that" is one way of saying to a friend devoted to the work of Damien Hirst that one's reluctance to join her on a visit to an exhibition featuring a dead cow, split in two, suspended in plastic, is in no sense an evaluation of her taste, or anyone's. In such cases, one wants, or doesn't want, certain things, and one's second-order motivational structure backs this up – one doesn't want to want otherwise. But one does not mistake this for valuing since valuing, we think, makes a stronger claim on behalf of what we aim at – a stronger claim upon others, and a stronger claim upon ourselves.

Lewis might say in his defense, that his proposal is simple, and "near enough" to capturing valuing. After posing the question, "What is 'valuing'?", Lewis writes:

> It is some sort of mental state, directed at that which is valued. It might be a feeling, or a belief, or a desire. (Or a combination of these; or something that is two or three of them at once; or some fourth thing. But let us set these hypotheses aside, and hope to get by with something simpler.) (114)

This is good philosophical practice, but note that one's second-order desires are themselves fairly "complicated" (116) and self-referential as a way of capturing a seemingly first-order, object-oriented psychological state like valuing. Think, for example, of the way a quite young child can *treasure* a doll or teddy bear, or *love* his parents or the family dog. What makes this attitude different from desiring to have these objects, or desiring to desire this?

Lewis gives us two other choices for the basic attitude involved in valuing: belief and feeling. We have already seen one of his grounds for rejecting belief – *believing to be valuable* is too close to

circularity to generate much philosophical insight. More importantly, he argues, if valuing is belief, in what sense is it a form of *favoring* (115)? On standard views of belief, belief that *p* – say, that the tide is now ebbing – is equally compatible with pro- or con-attitude relative to *p*, or with simple indifference to *p*. By contrast, on most views of valuing, including Lewis's, valuing is a pro-attitude. Lewis observes how we go about answering the question, "Do you value *X*?" We consult our intrinsic motivational dispositions; if unsure what these are, we attempt to simulate *X* and see whether we value it. Belief, on standard views, can function as it should in response to evidence of truth or falsity only if no such essential connection obtains – contrary evidence should reduce belief without reference to whether this is good or bad news for the believer, or whether this is what he *wants* to see happen. Citing previous work (Lewis 1988, 1990), he claims that attempts to locate a distinctive kind of content of belief which is such that the relevant belief that has inherent positive motivational force – turning it into what has come to be called a "besire" – run afoul of formal requirements on a rational updating function. The arguments here are complex, and ultimately inconclusive – there are simply too many ways to salvage the belief hypothesis. But it is safe to say that Lewis's reasons for rejecting belief as the analysans for value are strong indeed.

By contrast, Lewis's explanation of the unsuitability of *feeling* to analyze valuing is over in a wink. Here it is, in its entirety. Is value:

> A feeling? – Evidently not, because the feelings we have when we value things are too diverse. (114)

But surely the same argument, if it were good, would work against *desire*, or *desiring to desire*. We have a very great diversity of desires, too, when we value things – intrinsic desires, instrumental desires, *de dicto* desires, *de se* desires, cravings, goals, preferences, aversions, desires to do, to have, to protect, to pay, to reflect upon, to share with others, and so on. Within this bewildering array there are intrinsic desires to desire, and, by fixing upon these, Lewis attempts to identify the core of valuing. Why cannot we similarly fix upon certain feelings among the many?

Introducing feelings need not threaten the naturalistic reduction, since feelings such as "affection," "attraction," "interest," "anger," "fear," "liking," "happiness," and "sorrow" need not involve any normative concept. Perhaps there is no *one* basic feeling of "valuing *X*" – we *admire*, *appreciate*, *love*, *care about*, *resent*, *gloat*, *mourn*, *pine*, etc. But the core of valuing *X* might nonetheless be explained as an *affective* rather than strictly *motivational* state, one which involves dispositions to a particular pattern of feeling.

The psychologist's term for "feeling" is *affect*, and affect is well suited for the job description of valuing because this psychological kind is picked out by its role in systematically coordinating a range of responses. For example fear responds to contextual information by reorienting in a risk-relevant way attention, association, inference, memory, motivation, conscious feeling, physiological arousal, and action-readiness. Trust, by contrast, responds to evidence of safety or reliability by orienting these same processes in a way that is conducive to approach, cooperation, and acceptance. For this reason affective processes are called by psychologists *evaluative*, even when these processes often operate implicitly and involve no act of judgment (Schwartz and Clore 1983, 2003).

Valuing also influences in a systematic way one's responses to the world. What one values influences what one is typically disposed to want, and want to want, but also what one feels *affection* for, *protective* of, *happiness* in, and *sorrow* or *anger* about, as well as a host of more complex attitudes such as *approval* and *disapproval*, *admiration*, *pride*, *resentment*, *shame*, *guilt*, etc. A purely motivational state offers no explanation of why valuing has such a coordinated, pervasive effect on one's psyche, and why this palate of feelings is involved, rather than just *frustration* when the desire or desire to desire is not met, or *satisfaction* when it is. Valuing does bring with it a liability to feel frustration or satisfaction, depending upon how things go, but it also brings a good deal more.

We would surely question whether Ortcutt really values his children or family if he simply *desired to desire* to have offspring, but never noticed whether his children were disturbed or happy, never felt pain at their sorrow or pleasure in their happiness, never worried about their future or felt regret after realizing that he had let them down, never felt anger when they were treated badly, never found intrinsic enjoyment in spending time with them, and never felt sorrow, rather than frustration, when they abandoned him or died.

Would rejecting the desiring-to-desire analysis of valuing threaten Lewis's theory of value as a whole? Not fundamentally, I think. He writes, of the secondary-quality analogy for value: "…the analogy is none too good, and I doubt that it improves our understanding of either colour or value" (124). We might say something similar of Lewis's own starting point in dispositions to value. Recall the sequence of steps we rehearsed, above. One could start with a paradigm secondary-quality theory, looking something like this:

(3) *X* is red $=_{\text{def}}$ *x* is disposed to produce a red sensation in normal human observers when attended to visually under standard conditions.

Shifting to value:

(4) *X* is a value $=_{\text{def}}$ we are disposed to value *X* when we imagine *X* under ideal conditions.

And supplying Lewis's analysis of valuing, we get:

(5) *A* values $X =_{\text{def}}$ *A* is disposed to desire to desire *X* intrinsically.

And supplying his account of "ideal conditions", we arrive at:

(6) *X* is a value $=_{\text{def}}$ we are disposed to desire to desire *X* intrinsically under conditions of fullest possible imaginative acquaintance with *X*.

The psychological state of being disposed to value, as well as the concept of value itself, has disappeared from the right-hand side of (6). Did we ever need it? Suppose that (4) and (5) are false. Might (6) nonetheless be plausible in its own right? The fact that a life of lively, self-directed, challenging activity is something we are disposed to desire to desire for its own sake under conditions of fullest possible imaginative acquaintance, seems to be a reasonable ground for saying that living such a life would be of value to us, even leaving (4) and (5) entirely out of the picture. Dropping (4), modeled on (3), we complete the divorce Lewis rightly encourages from the secondary-quality picture. Dropping (5) opens the way for a more plausible Humean account of the state of valuing. Indeed, it opens the way for *Hume*'s account. For Hume analyzes evaluation not in terms of motivation as such, but in terms of *sentiment*, an attitude of approbation and disapprobation that can induce motivation.

34.4.2 Relativism and Realism

Lewis reduces facts about value to facts about our psychology, but these facts could have been different had evolution gone differently.

> Psychology is contingent. Our dispositions to value things might have been otherwise than they actually are. We might have been disposed, under ideal conditions, to value seasickness and petty sleaze above all else. Does the dispositional theory imply that, had we been so disposed, those things would have been values? That seems wrong. (132)

Might "seasickness and petty sleaze" have been values? Lewis indeed seeks to escape this conclusion, even though he admits that his theory "comes too close for comfort" to implying this (113).

How to avoid it? Lewis first explores two possibilities. First, one might "rigidify" the dispositions upon which values supervene (cf. Wiggins 1991). Defining value as what *actual* normal humans are disposed to value under ideal conditions might safely exclude seasickness and petty sleaze.[4] But Lewis argues that this is no solution – it merely moves the bump of contingency elsewhere under the carpet. For it would make it the case that, if we humans had happened actually to have dispositions to value seasickness and petty sleaze – a seeming possibility, albeit an improbable one – and had fixed the reference of 'value' in just the same way we do, as "that which we actually normally value under ideal conditions," then:

> In one good sense – though not the only sense – we would have meant by 'value' just what we actually do. And it would have been true for us to say 'seasickness and petty sleaze are values.' (133)

So Lewis considers a second, more direct way to avoid contingency – change the definition of value to include only that which we are *necessarily* disposed to value. Lewis canvasses several suggestions concerning the sort of necessity this would plausibly involve, fixes upon metaphysical necessity, and then argues, as we saw earlier, that the fire is worse than the frying pan: the cost of requiring necessity is bound to be an error theory, given the contingency of human dispositions to value (134). But life is too full of the wonders and horrors to think that nothing is really of value.

Therefore Lewis thinks the proper philosophical balance is to learn to live with the contingency of value, at least conditionally. He'll take "absolute value" – what all humans, or perhaps even all rational beings, are necessarily disposed to value under ideal conditions – if he can get it. But if he can't, he'll take "relative value," where the self-locating 'we' never gets replaced by 'all humans' or 'all rational beings'. This doctrine he calls "conditional relativity" about value (133), and he thinks that the contingency this involves is already tacitly known by most people (135).

It would undermine a noteworthy feature of evaluative practice – the giving and taking of information about value – if the scope of 'we' were, in practice, always just "me and those who think like me." But we know that those of different minds can nonetheless be disposed under suitable conditions to value many of the same things – including health and interpersonal decency – so relativism never rears its unattractive head in most daily life. We seldom have to worry that only a contingent bit of history explains this fact.

However, when it comes to dealing with those in very different situations or cultures, or those in our midst with fundamentally different mentalities (Lewis mentions colleagues in a philosophy department, without names), the relativism can come to the surface. If this pulls the rug out from under the practice of "browbeating" others to try to impose our own valuings upon them, so much the better (135).

That seems sensible, but I am not sure I see why the view should be developed in a *relativist* rather than *relationalist* (or *contextualist*) direction. "Relativism" is a much-contested notion, so there is no proprietary sense, but it seems to me that one serviceable core sense is something like this: we can speak of a magnitude or property as *relativistic* if it is indexed to the *evaluator's* point of view. Consider a life of high philosophical engagement and endless discussion. This might be the sort of thing some of us are disposed to value, while others are disposed to disvalue heartily. It also might be the sort of life that *suits* some people well and others not at all.

If I am a relativist, the question whether such a life is a value will be answered by focusing on my dispositions to value it under ever fuller conditions of full imaginative acquaintance. The 'we' in (6) is centered on me. With luck, I might find others of the same mind, but that is not part of what is required for me to make the judgment.

If I am a relationalist, the question whether such a life is a value will be answered asking whether it would suit whoever would be living it, which might or might not be me. So we'd have something more like:

(7) *X* is a value for $A =_{\text{def}}$ *A* is disposed to desire to desire *X* intrinsically under conditions of fullest possible imaginative acquaintance with *X*.

Relationalism is not absolutism – it does not presuppose that one set of values fits all. But it is not relativism, either. For there is no standpoint dependence in (7) – once *A* is fixed, it should yield the same answer *whoever* is doing the evaluating.

Lewis at one point speaks of *good*:

> I don't say that everything we value is good; but I do . . . say that to be *valued* by us means to be that which we desire to desire. Then to be a value – to be good, near enough – means to be that which we are disposed, under ideal conditions, to desire to desire. (116)

The relationalist would be saying, in effect:

(8) *X* is a good for $A =_{\text{def}}$ *A* is disposed to desire to desire *X* intrinsically under conditions of fullest possible imaginative acquaintance with *X*.

Good for *A* is different from absolute good or good *simpliciter*. We might speak of a *conditional relationalism*, then, on the model of Lewis's *conditional relativism*. One might start off with the hope that all humans, or all rational beings, would share a core set of second-order desires under ideal conditions. But (8) does not tip us over into nihilism if this is not the case – it permits, but does not require, convergence for all *A*. Moreover, (8) might be thought to preserve a form of existence internalism more plausible than (6), since it ties what is good for an individual to what could engage *that* individual, rather than what could engage the evaluator. (8) might also support a more robust pluralism than (6), and thus answer better to Lewis's underlying values. For (8) makes room for me to accept, in my own voice, the conclusion that there are, or can be, genuinely good lives for others that, given my capacities, or mentality, or society, or culture, simply would not be an object of my second-order desires with respect to my own life. An alarming prospect! – or exhilarating, to those of us who delight in the rich variety of life.[5]

The difference between conditional relativism and conditional relationalism is large, I think, though each is compatible with the broad grounds Lewis invokes in "Dispositional Theories of Value." Conditional relationalism would have a better fit, I think, with his underlying pluralism, which is tied to a lively sense of the delightful and devilish variety among humankind. Good for *A* varies with *A*'s nature, not who's judging. Moreover, it would have the advantage for Lewis of making it clearer why his view might in some measure deserve the name of *realism* (132ff.). Good for *A* is *standpoint-independent*, just like nutrition for *A*. Relatedly, relationalism might help assuage some of the concern about contingency. For we can speak of value as a *constant function* for each *A*, even though it yields different values for different *A*s. '$X + 2$', for example, is a constant function: its value depends upon the value of *X*, but not upon who's asking. 'That which *A* would desire to desire under fullest imaginative acquaintance' is likewise invariant, despite yielding different answers for different values of *A*. And there need be nothing contingent about *that*.

34.4.3 Analyticity

Lewis offers his account as a "non-obvious" analytic truth, which, moreover, is "equivocal" given the many ways in which the term 'value' is used (129ff).

This is consistent with his larger project of making the world safe for analytic truth – in science, in the philosophy of mind, in metaphysical inquiry. We needn't here enter into the merits of this project in a general. But for the case of value, I think, it is especially problematic to offer (6) as an analytic definition – as opposed, perhaps, to a reforming definition or synthetic property identity.

Consider the moral of Moore's Open Question argument, or, for that matter, of Hume's "*is*/*ought*" argument, or Kant's "analytic/synthetic" distinction. According to (6):

(6) *X* is a value $=_{\text{def}}$ we are disposed to desire to desire *X* intrinsically under conditions of fullest possible imaginative acquaintance with *X*.

Moore might counter: "Consider Lewis's discussion about whether (6) should be modified to:

(6N) *X* is a value $=_{\text{def}}$ necessarily, we are disposed to desire to desire *x* intrinsically under conditions of full imaginative acquaintance with *X*

Lewis rejects (6N) on grounds of costs and benefits – in particular, that (6N) seems bound to 'bring on error theory' about value, but nihilism is simply not credible. But in saying this Lewis is making a substantive normative claim – synthetic in nature, and not something a *definition* could settle."

We see another example of a substantive evaluative judgment at work when Lewis writes, concerning why he restricts the information involved in assessing value to vivid simulation:

> Another unhelpful sort of knowledge is a vivid awareness that we are small and the cosmos is large; or a vivid awareness of the mortality of all mankind, and of the cosmos itself. If such knowledge tends to extinguish all desire, and therefore all valuing, it will not help us to value just what is valuable. (124–5)

But philosophers in various cultural traditions might say that such extinction of desire is just what is needed to give us the perspective to see what is truly valuable. And philosophers in more proximate cultural traditions might say that cosmic information has an important tempering effect on egocentric concern: we should see ourselves as just one individual among many, one species among many, and one planet among many. For example, Bertrand Russell urged such humility, writing in *What I Believe* that "the earth is merely one of the smaller planets of one of the smaller stars in the Milky way" (Russell 1925). Frank Ramsey wrote in reply,

> If I was to write a *Weltanschauung* I should not call it "What I believe" but "What I feel"…. Where I seem to differ from some of my friends is in attaching little physical importance to physical size. I don't feel the least humble before the vastness of the heavens. The stars may be large, but they cannot think or love; and these are qualities which impress me far more than size does. I take no credit for weighing nearly seventeen stone. (Ramsey 1931, 248–9)

It perhaps is no surprise that Ramsey and Lewis would share such feelings about the irrelevance of cosmic vastness to value, but it is certainly clear in Ramsey that this is a substantive evaluative attitude, best understood as a "feeling" of what is or is not important, and not something that we can settle in giving the meaning of 'value'. Indeed, we need the word 'value' to capture the bone of contention between Ramsey (or Lewis) and Russell.

(6), a putatively *a priori* analytic definition, might therefore be reformulated as a synthetic proposal about the nature of value – an *a posteriori* property identity:

(8) *X* is a value = *X* is such that we are disposed to desire to desire *X* intrinsically under conditions of fullest possible imaginative acquaintance with *X*.

Understood in this way, Lewis's account might indeed offer the best balance of respecting ordinary intuitions, familiar platitudes, practices of evaluation, standards of clarity, and paradigm cases. To defend it, however, will require defending substantive normative judgments – defending subjectivism over objectivism, relativism over relationalism, and the irrelevance to the value *X* of factual information beyond the deliverances of the fullest possible imaginative simulation of *X*.

Ramsey also wrote, despite the ill health that cut short his life:

> Humanity, which fills the foreground of my picture, I find interesting and on the whole admirable. I find, just now at least, the world a pleasant and exciting place. (1931, 249)

Lewis, I think, could have penned much the same words. Each boldly embraced life in the face of unkind fate. And each gave us the best of reasons for doing so – teaching us, by example, just how interesting and admirable humanity can be.

Notes

1 Lewis 1989, all unattributed page references in the text are to this essay.
2 Lewis notes a qualification for ways in which imaginative acquaintance might alter his dispositions in the case of "finkish" dispositions to value (117, n.).
3 The qualification 'some' is important here. Take the case of prescriptivism. Lewis appears to have in mind theories akin to Ayer's emotivism, in which it is enough – first, last, and always – for me sincerely to assert that *x* is good that I have a motivating attitude toward *X* (Ayer 1930).
4 'Sleaze' was a favorite term of Lewis's. Since the term struck me as unusual, I once looked it up. The explanation of its origin surprised me and I reported it to Lewis, who was always interested in linguistic *arcane*. Apparently, 'sleaze' and 'sleazy' derive from an ethnic slur. Nineteenth-century Brits, distressed at the flood of cheap textiles from imported Eastern Europe, which threatened British dominance of the mass textile market, spoke disparagingly of the (putatively) poor quality imported cloth as "sleazy" – deriving from 'Silesia'. The term then generalized and took on new meanings. Rather than thinking this long-forgotten fact a reason to avoid the term, however, Lewis found it delightful. He was always glad, he said, when ethnic slurs lose their ethnicity, and with it their power to hurt. I, in turn, was pleased to see the term show up in "Dispositional Theories of Value."
5 For further discussion of relationalism vs. relativism, see Railton 1986a, 1986b.

References

Ayer, A.J. (1930). *Language, truth, and logic*. New York: Dover.
D'Arms, J. & Jacobson, D. (2000). Sentiment and value. *Ethics*, 110, 722–48.
Darwall, S. (1997). Reasons, motives, and the demands of reality: an introduction. In S. Darwall, A. Gibbard & P. Railton (Eds.), *Moral discourse and practice*. Oxford: Oxford University Press.
Hume, D. (1985). *Of the standard of taste*. In *David Hume: essays moral, political, and literary*, pp. 231–57. Ed. E.F. Miller. Indianapolis: Liberty Fund. (Orig. work 1760.)
Lewis, D. (1980). A subjectivist's guide to objective chance. In R. Jeffrey (Ed.), *Studies in inductive logic and probability*, pp. 263–93. Berkeley: University of California Press.
Lewis, D. (1986). Postscript to "A subjectivist's guide to objective chance." In D. Lewis, *Philosophical papers*, Vol. II, pp. 114–32. Oxford: Oxford University Press.
Lewis, D. (1988). Desire as belief. *Mind*, 97, 323–32.
Lewis, D. (1989). Dispositional theories of value. *Proceedings of the Aristotelian Society*, suppl. vol., 113–37.
Lewis, D. (1996). Desire as belief, II. *Mind*, 105, 303–13.

Mill, J.S. (1962). *Utilitarianism*. Reprinted in M. Warnock (Ed.), *John Stuart Mill: utilitarianism and other writings*. New York: Penguin. (Orig. work 1863.)
Railton, P. (1986a). Moral realism. *Philosophical Review*, 95, 163–207.
Railton, P. (1986b). Facts and values. *Philosophical Topics*, 14, 5–31.
Ramsey, F. R. (1931). *Foundations of mathematics and other logical essays*. Ed. R.B. Braithwaite. London: Trench, Tubner.
Russell, B. (1925). *What I believe*. New York: Dutton.
Schwarz, N. & Clore, G.L. (1983). Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. *Journal of Personality and Social Research*, 45, 513–23.
Schwarz, N. & Clore, G.L. (2003). Mood as information: 20 years later. *Psychological Inquiry*, 14, 296–303.
Wiggins, D. (1991). A sensible subjectivism? In D. Wiggins, *Needs, values, truth*. Oxford: Blackwell.

35

David Lewis's Social and Political Philosophy

SIMON KELLER

35.1 Introduction

David Lewis wrote seven papers on topics squarely within the fields of social and political philosophy. Two are about toleration, two are about deterrence, two are about punishment, and one (which is very short) is about obligations to the distant poor.[1] The papers do not receive as much attention as does Lewis's work in other fields, and only one of them – "Mill and Milquetoast," on John Stuart Mill's defense of toleration – is prominent in the scholarly literature on its topic.[2] One reason for this, probably, is that the papers tend to concern problems marginal to mainstream social and political philosophy. Another, probably, is that that the papers do not themselves make extensive reference to the extant literature. The papers are probably better appreciated by fans of David Lewis than by philosophers working in the fields on which they bear.[3] Whatever the case, the papers are all entertaining and insightful, and they reveal something about Lewis's broader approach to philosophy, as well as having much philosophical value in their own rights.

I shall consider in turn Lewis's views about toleration, deterrence, punishment, and obligations to the distant poor, and then I shall ask what overall perspective in social and political philosophy we might take him to hold. I will try to make Lewis's views clear and to emphasize points suggestive of his overall perspective, and I will also offer some critical comments. In my opinion, Lewis's major claim about toleration does not take him as far as he thinks, and his major suggestion about punishment does not ultimately succeed on its own terms. In the relevant sections, I shall try to explain why.

35.2 Toleration

35.2.1 Neutralist Justifications of Toleration

Here in our liberal societies, we value toleration. With very few exceptions, we think people should be able to hold and express their opinions, even if they are wrong, even if they are dangerous. But if

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

an opinion is wrong, and if an opinion is dangerous, then what is the good in tolerating it? Why have institutions of toleration, rather than institutions that advance truth and discourage falsehood? Lewis addresses the general question in "Mill and Milquetoast," and its application to a special case in "Academic Appointments: Why Ignore the Advantage of Being Right?"

The case studied in "Academic Appointments" is of a philosophy department deciding whom to appoint to a vacant position. According to standards of academic decorum, Lewis reports, it is not acceptable to support a candidate on the ground that her philosophical views are true, or to reject her on the ground that her philosophical views are false (190–1). In this respect, all philosophical views are tolerated. But if the point of philosophy is to uncover the truth, then why should the truth of a candidate's views not be a legitimate consideration? Are we not indeed obliged to hire philosophers whose research takes them closest to the truth (189–92)?

A number of justifications for toleration come quickly to mind. We can never be certain that our own opinions are true, so we can never be certain that in suppressing what we see as falsehood we do not in fact suppress truth. Truth and falsity can be entangled, so in suppressing a falsehood we may also suppress some truth with which it is linked. It is good for true opinions to be challenged, so that they do not become dogma; it is good, for example, to have some wrong-headed philosophers around to keep the right-minded on their toes. There are costs to suppressing speech, and a society in which falsity is tolerated may be better overall than a society armed with weapons of suppression.

These justifications of toleration are all "neutralist" justifications, to use Lewis's term, because they do not presuppose any of the particular opinions at issue (161). People of all opinions can see, for example, that there are benefits to having conflicting opinions represented in a society, and in a philosophy department. There are other neutralist considerations in support of toleration; Lewis lists many more (164–7, 192–7). And Mill's project in *On Liberty*, Lewis says, is to justify toleration by appeal only to neutralist justifications (161–3). But Lewis doubts that neutralist justifications can fully explain why we should tolerate false opinions, for three reasons.

First, neutralist justifications tend to depend on overly ambitious empirical claims (164–7, 192–7). Can we really never be sure that some opinion is false? If we tolerate false opinions, will true opinions really come to be more vigorously defended and better understood, or will they instead be swamped and forgotten?

Second, neutralist justifications do not seem to show that the truth or falsity of opinions should be ignored (196–7). There may be benefits to tolerating falsity, but there are also benefits to promulgating truth. Should we not weigh these benefits against each other, rather than treating truth and falsity as irrelevant? Should not the correctness of a candidate's philosophical views, for example, at least count for something?

Third, we can imagine people for whom the neutralist arguments would have insufficient force, and among them are some of the people we most need to convince. In "Mill and Milquetoast," Lewis gives the example of the Inquisitor (167–73). The Inquisitor accepts the neutralist arguments, as far as they go, but also believes that heretics suffer eternal damnation and that if we tolerate heresy then it will spread. Whatever the benefits of tolerating heresy, says the Inquisitor, they are outweighed by the harms. Lewis does not give a parallel case for academic appointments, but I suppose we could imagine a convinced naturalist, who thinks that non-naturalist programs in philosophy are not only false but also seductive to weak minds and an embarrassment to the discipline. Non-naturalism in philosophy departments might serve some purpose, our naturalist would say, but not so great a purpose as to counter its dangers.

35.2.2 Treaties of Toleration

Lewis thinks a better defense of toleration is available. His suggestion is that people of different opinions can all have good reasons, though not the same reasons, to agree to a tacit treaty on which all opinions are tolerated.

Imagine an atheist who thinks Christianity false and dangerous. What she would most like is for Christianity to disappear, and what she would least like is for the Christians to take power and impose their religion on everyone. Imagine a Christian with analogous but opposite preferences. He would like atheism just to go away, and thinks it would be terrible if atheism were to triumph and Christianity to be suppressed. Neither atheism nor Christianity is inclined to depart voluntarily, so if the atheist is to suppress Christianity, or if the Christian is to suppress atheism, then there must be a war, literal or metaphorical. In a war, each side faces the possibility of defeat. The atheist might get she what wants most, but she might get what she wants least; so too for the Christian.

It may then serve the interests of the atheist, and also the different interests of the Christian, to agree not to have a war. They could make a treaty on which neither will try to suppress the opinion of the other. Then they each give up the quest to attain their respective most preferred outcomes, but in return they each avoid the costs of going to war, and receive a guarantee that they will not face their respective least preferred outcomes. With the terms of the treaty set, each party may indeed think quietly that it has secured the better deal (178–80, 198–9). Under fair conditions, the atheist may think, the good arguments for atheism will win more minds than the deceitful arguments for Christianity. If it comes down to the merits of the case, the Christian may be sure, then the advantage goes to Christianity: the view with truth on its side.

In reality, whether you are the atheist or the Christian or neither, there will be more than one person who finds your opinions about religion false and dangerous, and you will have many contentious opinions apart from those concerning religion. So in reality, you will need not just one treaty with one religious adversary, but many treaties with many different adversaries on many different disputes. Obtaining such a multitude of treaties on prudent grounds will be difficult, especially when we add that the facts about who agrees and who disagrees with you, and about how power and fanaticism are spread between your various allies and adversaries, will differ across different disputes and change constantly over time. It would suit you better if there were one big treaty, on which everyone agrees to tolerate all the opinions of everyone else. For their own separate but analogous reasons, others will prefer to have the big treaty too (182–4).

A similar story can be told about academic appointments. You may find certain philosophical views false and dangerous, but others will think the same of your views, and those others may one day be in the majority. In a discipline featuring many different opinions, and characterized by trends and changes in received wisdom, one big treaty will offer more security than many small treaties. For you, a treaty on which everyone ignores the truth and falsity of job candidates' views offers the best available protection of the true philosophical views. The same is true for other philosophers, even if they do not agree with you about which philosophical views are the true ones (199–200).

Institutions of toleration are then justified, for Lewis, because they serve a tacit treaty of toleration, which everyone has some reason or other to accept. All parties can agree that the treaty offers the greatest available balance of benefits over harms, even though they do not agree about which are the benefits and which are the harms (199). In this sense, the treaty is "a contract for utilitarians" (179).

Lewis's justification of toleration does not require us to accept the empirical claims lying behind neutralist justifications of toleration; nobody needs to believe, for example, that things are better if false views prosper. The justification gives a reason why the truth and falsity of opinions should be

ignored, not just balanced against other considerations; its point is to prevent anyone from imposing her own opinions on the discussion. Further, the justification is designed explicitly to appeal to a character like the Inquisitor, who sees the most enormous of harms in opposing opinions; it directs him to compare not toleration with suppression, but the certainties of toleration with the uncertainties of war.

35.2.3 The Force of a Tacit Treaty

I think that Lewis's justification of toleration depends upon a doubtful empirical supposition of its own, and that this leaves it open to criticisms just like those that Lewis directs at neutralist justifications.

Imagine an updated version of the Inquisitor, who is convinced by Lewis's argument but occasionally sees opportunities to suppress dangerous falsehoods without facing repercussions. He might think, "If I move quickly, then I can stamp out this latest resurgence of liberation theology before it gathers momentum, and if I am sufficiently discreet then nobody much will notice. Our institutions of religious toleration will survive and the truth will still be protected." (Or, imagine a version of the philosophical naturalist, who thinks, "This job candidate, so clever and charismatic, with his new non-naturalist theory of meaning, is all set to do some real damage. With some strategic voting I can keep him out, and this one move is hardly going to undermine the treaty of toleration from which the true naturalist philosophy benefits. I would be crazy not to strike.") What can we say about intolerance in cases like this?

The challenge is not to show that nobody would in fact be intolerant under such circumstances. Some people are intolerant when they think they can get away with it, and no view about the ethics of toleration is going to change that. The challenge is to show why, in such cases, intolerance is yet wrong.

We cannot say that the Inquisitor's intolerance is wrong because it constitutes a failure to honor a commitment. The treaty of toleration is a tacit treaty, not a real treaty. Nobody actually signs a tacit treaty. A tacit treaty is less like a real treaty than it is like a convention. If you and I formally agree to cease hostilities, then, all else equal, you do something wrong if you do not keep to the agreement: you break your promise. Someone who breaks a convention, on the other hand – someone who eats her main course with her dessert spoon, for example – need not thereby fail to do something she said she would do, and this is so even if the convention serves her interests and she has no wish to undermine it. Whatever else might be wrong with the updated Inquisitor, he need not be guilty of breaking a promise.

We could tell the Inquisitor that he is being imprudent. Perhaps he would be better off adopting a policy of never suppressing others' opinions, because he can never be sure that his actions will not redound to the harm of the truth (as he sees it). But that is a hopeful empirical assumption, at best, and one that a reasonable person could sometimes reject.

There are other criticisms available. When we benefit from an institution of toleration, perhaps we acquire a moral obligation to conform to its rules. When we suppress opinions while enjoying other people's toleration of ours, perhaps we take advantage of other people, or treat them as mere means. If everybody suppressed opposing opinions when they could, then institutions of toleration would fall apart; perhaps that is a reason why it is wrong for even one person to engage in suppression, even when she knows that others in fact will not. Those criticisms may be well founded, but they take us outside the framework of the "contract for utilitarians." They seek to give us reasons to tolerate falsity even when doing so will not help protect (what we see as) the truth.

Lewis's justification of toleration has force in some cases in which neutralist justifications do not, but, like the neutralist justifications, it depends upon an empirical supposition that will sometimes

fail to hold, and it is at its weakest, normatively and motivationally, when the most cunning and fanatical see opportunities to suppress the opinions of the most vulnerable.

35.3 Deterrence

35.3.1 The Paradox of Deterrence

Lewis's papers on deterrence are "Devil's Bargains and the Real World" and "Buy like a MADman, Use like a NUT." The papers are concerned, each in a different way, with the paradox of deterrence. The paradox of deterrence arises when you have good reason to make a sincere threat, but also good reason not to act on that threat should the occasion for acting on it arrive (201).

Lewis focuses upon the case of nuclear deterrence (205, 219–23). Our enemy has nuclear weapons, which will inflict terrible damage upon us if used. If we threaten to meet a nuclear attack with nuclear retaliation, then we give the enemy a reason not to attack. We can make our threat convincing, however, only if we truly intend to follow through on it; perhaps we must set up the procedures required for immediate retaliation and give the appropriate conditional orders to the military. But, if the enemy does attack, we will not have good reason to retaliate. Retaliation will just mean more innocent deaths and an escalation in hostilities. So we seem to have good reason to intend to retaliate if attacked, but also good reason not to retaliate if attacked.

There are other cases in which the paradox of deterrence arises. Perhaps the best way to make your students respect deadlines is to convince them that you will refuse to grade any papers that are not submitted on time. Perhaps your students are very good at telling when you are serious and when you are not, and will only be convinced by your threat if you deliver it sincerely. And, perhaps, if a student does submit after the deadline, then the right thing to do is to grade her paper anyway; to refuse to grade the paper, at that point, would be pointless and heartless. So, you might have reason to intend to refuse to grade late papers, but also reason to grade all papers, even those that come in late.

In "Devil's Bargains," Lewis considers and rejects two attempts to resolve the paradox, and then offers a solution of his own. The first attempted resolution says that it is wrong to intend to do something you ought not to do; we should not intend to do such a horrible thing as to unleash pointless nuclear retaliation. The second attempted resolution says that it is permissible to act on an intention you have good reason to form; we may unleash pointless nuclear retaliation if attacked, because there is such good reason to form an intention to do so (201–4).

Lewis's solution is to embrace both sides of paradox and wonder why this should be a problem (202–4). If the question is whether we should form the intentions that deter undesirable behavior, then Lewis's answer is yes, we should, because they serve a good purpose. If the question is whether we should perform the actions we thereby intend to perform, then Lewis's answer is no, we should not, because no good will come of it.

The advantage of Lewis's solution is that it yields plausible verdicts on the intentions and the actions, considered for their own harms and benefits. Its possible disadvantage is that it does not help us to evaluate the persons who display those intentions or perform those acts (205–7). We might want to know, for example, how to judge a leader who forms the intention that most effectively protects her country, but thereby stands ready to unleash pointless indiscriminate destruction. We might want to know how to judge a teacher who is always sensible and compassionate in dealing with late papers, but only because she fails to embrace the policy that would teach her students to respect deadlines.

Lewis is happy to say in such cases that no overall judgments of persons can be found (205–7). We can evaluate intentions and failures to intend, and acts and failures to act, but we may not always

be able to evaluate persons. Overall judgments of persons are needed by a god on Judgment Day, deciding who to save and who to damn, but Lewis, he tells us, does not believe in such a god, so that is no concern for him (207–8).

A consequence of Lewis's solution is that it is not always possible for a person to be fully virtuous, at least on the plausible view that a fully virtuous person does not intend to do horrible things. There are some predicaments, Lewis's solution implies, that cannot be negotiated without either a failure to form the best intentions or a failure to intend to perform only the best actions. Accordingly, Lewis's solution could not be offered by a paradigmatic virtue ethicist: one who evaluates acts and intentions by asking whether they would be performed or displayed by a fully virtuous person.[4] The tone of Lewis's paper is such as to suggest that he would not find these implications troubling. Too bad, he would likely say, for the supposition that full virtue is always possible, and too bad for virtue ethics.

35.3.2 Existential Deterrence

The main argument of "Devil's Bargains" takes it for granted that in the cases under discussion, it is possible to form an intention in a manner that allows it to act as a deterrent. But that might not be so easy. We can say that we will retaliate if attacked, but if the enemy knows (and knows that we know) that there will be no point in our retaliating once an attack occurs, then why should he take our threat seriously? This is the question taken up in "Buy like a MADman," which considers two strategies for nuclear deterrence.[5]

The first strategy is Mutually Assured Destruction: MAD, advocated by the MADman. The MADman would have us build a nuclear stockpile that enables us to inflict damage so horrendous that the enemy will rationally judge that even if we are unlikely to retaliate to an attack, the risk that we will is not worth taking. The MADman's strategy requires us to prepare to unleash horrible indiscriminate destruction, but it has the advantage, in the nuclear age, of being relatively inexpensive. We don't need to have the best nuclear weaponry available, and we don't need to get into an arms race. We only need to be able to destroy the world once (220–1).

The second strategy comes from the Nuclear Use Theorist: the NUT. The NUT asks for a plan of nuclear retaliation that will damage the enemy while also serving the military ends we will have at the time of retaliating: a plan for fighting a nuclear war. The NUT's strategy generates a believable threat and does not require us to prepare to unleash horrible indiscriminate destruction. But it is expensive, and it is likely to lead us into an arms race, because it requires us always to possess the weapons needed to fight the enemy and win. And, it increases the risk of nuclear war, because it creates an incentive to attack preemptively. If there is to be a war, then each side would prefer to be the one who gets things started by destroying some of the other side's capability (221–3).

Lewis thinks we can take something from each strategy, in the service of a strategy called "existential deterrence." He gives the example of a burglar deciding whether to enter a house guarded by a tiger. Tigers are scary and unpredictable. However well the burglar may have prepared, and whatever he may think he knows about this particular tiger, his governing thought, if he is sensible, will be, "You don't tangle with tigers." In formulating our nuclear strategy, Lewis suggests, we should aim to be like a tiger, sufficiently scary and unpredictable to convince a sensible enemy – a "risk averse, pessimistic, skeptical, and conservative" enemy – that a fight with us is best avoided (223).

Lewis's direction for creating an existential deterrent is to "buy like a MADman," procuring just the weapons needed to inflict horrible destruction if the enemy attacks, but, if attacked, to "use like a NUT," employing those weapons in ways that protect us and hurt the enemy's capability while minimizing overall destruction (224). Part of the trick is to buy the kinds of weapons most useful to a NUT, but only to the level demanded by the MADman; we should procure weapons meant for fighting a nuclear war, but only as many of them as we need to unleash horrible indiscriminate

destruction, if we choose (225–6). That way, we keep down costs, avoid an arms race, and take away the incentive to strike preemptively. We give ourselves the option of responding to an attack with either strategic warfare or massive destructive vengeance. We leave the enemy sure that we are dangerous, and unsure what we will do if provoked.

As Lewis seems to acknowledge, we may create a more effective existential deterrent if we ourselves do not know for certain how we would respond to a nuclear attack (215–17). It will only enhance our scariness and unpredictability if we can honestly say, "If they attack us, it will make most sense for us to respond strategically. But in the circumstances following an attack, the facts on the ground will not be clear, and we will not know exactly what a good strategic response would be. And we will be distressed and angry. We might show restraint, or we might unleash horrible indiscriminate retaliation, just because we can. Who knows what we would do?" Full self-transparency is not necessarily an advantage, when you want to look like a tiger.

35.4 Punishment

35.4.1 Two Practices of Punishment

Lewis's papers on punishment are "Do We Believe in Penal Substitution?" and "The Punishment That Leaves Something to Chance." In each paper, Lewis notices something suspicious about an accepted practice of punishment, and then wonders whether that practice is defensible.

Penal substitution occurs when one person takes the punishment for a crime committed by another. In some cases, we appear to have no objection to penal substitution. If your punishment is to pay a fine, and I offer to pay the fine for you, then the law has no objection; all that matters is that the fine gets paid (133–4). Christian doctrine says that Jesus was able to pay for our sins through his suffering (130–1). In other cases, however, the idea of penal substitution strikes us as ridiculous – I cannot volunteer to serve your prison term – and it seems that the whole point of punishment, really, is that the wrongdoer herself face consequences for her actions (128–30). In "Do We Believe in Penal Substitution?" Lewis describes our divided attitude to penal substitution and paints it as a problem for all of us (not just for Christians). In closing the paper, however, he reiterates the point that we all seem to find penal substitution acceptable in some cases, and offers this as a reason to think that penal substitution is indeed sometimes acceptable – even if we cannot say why (135).

'The Punishment That Leaves Something to Chance" begins with our practice of punishing successful criminal acts more harshly than unsuccessful criminal acts. If you and I both set out to commit murder, and only I succeed, then my punishment for murder will be greater than your punishment for attempted murder. This will be so even if we act with equally evil intentions, and even if it is only through sheer luck that my act succeeds while yours fails. From the standpoint of justice, however, you would not seem to deserve a lesser punishment than me, just because you were not able to do the awful thing you tried to do. It would seem that agents who act with equal wickedness, or with equal "heartedness" (as in "wholehearted" and "halfhearted"), should suffer equal punishments (227–8). Does it follow that our practice is unjust?

Lewis canvasses some possible justifications for punishing successful attempts more harshly than unsuccessful attempts, and finds none of them satisfactory. He then offers an ingenious and elaborate suggestion of his own. The suggestion is offered ambivalently – Lewis says with emphasis "*I do not say that it works*" (232) – but he thinks it captures the real nature of our practice, and he cannot find anything wrong with it.

The crucial idea behind Lewis's justification is of a "penal lottery." If you want to reward a team for a good performance, you might give each member of a team a certain number of lottery tickets.[6]

Some of the tickets will win money and some will not, so some members of the team will ultimately receive greater benefits than others. To that extent, different members of the team get different rewards. There is another sense, however, in which the reward is the same for all. Each is rewarded with the same chance of winning money in the lottery.

A penal lottery works similarly. A person commits a wrong, and as punishment he receives a certain chance of suffering a certain kind of harm. Lewis gives as an example the practice of punishing the soldiers in a mutinous regiment by executing one soldier in every ten; each soldier's punishment is to get a one-in-ten chance of execution (233). Lewis is not sure whether giving someone a ticket in a penal lottery can be a just method of leveling punishment – this is the source of his ambivalence – but he can find no persuasive reason to think that it cannot (234–8). And he thinks that if we grant that a penal lottery can, in principle, be just, then we can find our rationale for punishing successful criminal attempts more harshly than criminal failures.

35.4.2 A Penal Lottery by Enactment

When a criminal sets out to commit a murder, she may or may not succeed. The chance that she succeeds is the same as the risk to which she puts the life of her intended victim. We can then think of the criminal as buying a ticket in a lottery. Depending on the outcome of the lottery, her attempt becomes either a murder or a mere attempted murder. The greater the risk she imposes on the life of the intended victim, the greater her chance of becoming a murderer.

Because we punish murder more harshly than mere attempted murder, the criminal becomes eligible for a harsher penalty if the lottery has one result rather than the other. Her act draws a harsher penalty if it becomes a murder. But that is just to say that if the criminal is caught and convicted, then the lottery in which she buys a ticket turns out to be a penal lottery. In imposing a certain risk upon the intended victim's life, the criminal takes on a certain chance of performing a murder rather than a mere attempted murder, and hence a certain chance of facing a harsher penalty rather than a milder penalty.

Our practice of giving harsher punishments for successful criminal attempts is then underpinned, says Lewis, by a penal lottery. It is not as neat a penal lottery as the simple one described earlier. For one thing, in this lottery, everyone with a ticket receives a penalty; the lottery determines whether the criminal faces a greater or a lesser harm, not whether she receives some harm or none. For another, the result of the lottery is determined before the criminal is convicted; by the time a case comes to trial, it is known whether the accused person will face the greater harm (the punishment for a successful attempt) or the lesser harm (the punishment for a failed attempt), if she is found guilty. But those complications do not appear to make a difference to whether the penal lottery is just or unjust. If a simple penal lottery is just, Lewis thinks, then so is this one (238–42).

Tickets in this penal lottery are not given for acts of murder, or for acts of (mere) attempted murder. Instead, the criminal gets a ticket for the crime of putting a person's life at risk. If you and I make the same attempt, so to speak, to kill someone, then we each impose the same risk on the person's life, so we each have the same chance of succeeding in our attempt, so we each get a ticket in the same penal lottery. If only one of us succeeds, then we do not ultimately receive the same penalty, but still, in one sense, for our acts of equal wickedness we get the same punishment: the same chance of suffering the harm meted out for a successful versus a failed attempt to murder.

So, Lewis concludes, we can avoid the charge that our practice gives unequal treatment to people who have equal deserts. Criminal acts of equal heartedness draw the same punishment. It is just that the punishment is a ticket in a penal lottery.

35.4.3 Objective Risk and Heartedness

I think that Lewis's solution does not quite perform as advertised. It uses the objective chance that a criminal attempt will succeed as a measure of the heartedness with which the criminal performs the attempt; a prospective murderer's chance of becoming a successful murderer is set by the risk to which she puts her victim's life (241–2). Objective chance and heartedness often come apart, however. Consider two cases.

You hear tell of a kind old hobo who lives behind a certain bush at a certain place in the desert, and you resolve to murder him. You ride into the desert, locate the bush, take careful aim, fire in a barrage of bullets, and ride back to town. I hear tell of another kind old hobo who lives behind a bush in the desert, and I too resolve to find him and kill him. I ride into the desert, find the bush, take careful aim, unload my shotgun, and ride back to town. As it happens, the story you hear about a kind old hobo in the desert is true, but the story I hear is entirely fabricated. You fire a barrage of bullets into a person's flimsy home. I fire a barrage of bullets into an unremarkable bush in an uninhabited desert, imagining it to be the home of someone who in fact has never existed. Your act puts a person's life at great risk, and gives you a very high chance of becoming a murderer. My act causes hardly any risk to anyone, and gives me hardly any chance of becoming anything other than a failed attempted murderer. Yet, we act with equal wickedness and resolve.

A second case: we each hear tell of a hobo who lives behind a bush in the desert, and we each ride into the desert with murderous intent. I take great care in identifying the correct bush, and I deliberately and gleefully fill it with bullets. You are much less firm in your resolve, and cannot find the motivation to do any research into the exact location of the hobo's home. Instead, you ride a little way into the desert, unload your shotgun into the first bush you see, and then ride home, wondering whether you got the right one. What neither of us knows is that many, many hobos live here in this desert; in this desert, in fact, every bush harbors a hobo. As it turns out, your act of shooting into a bush chosen at random is no less likely to kill a person than is my act of shooting into a bush chosen precisely because I know someone lives behind it. As it turns out, we each impose the same risk upon a person's life, and we each have the same chance of committing a murder as opposed to a mere attempted murder. Yet, my attempt to kill is wholehearted, while yours is halfhearted.

The objective chance that a criminal attempt succeeds can fail to reflect the level of heartedness with which the attempt is carried out. Whether a criminal attempt is likely to succeed will often depend upon factors that have nothing to do with the wickedness or resolve of the criminal. Lewis's penal lottery, then, will often fail to apportion likelihood of greater and lesser punishments to heartedness. Under Lewis's construal of our practice, we distribute tickets in a penal lottery – we levy punishment – according to the objective chance that a criminal act succeeds, not according to the degree of heartedness with which the act is performed. From the standpoint of justice, this leaves our practice looking as difficult to justify as ever. Put it this way. If we know that a failed criminal acted wholeheartedly, but also that his attempt was bound to fail, should we not enter him in a new lottery, in which he has a greater chance of earning a harsher punishment? And if we know that a successful criminal acted halfheartedly, but that it was almost inevitable in the circumstances that he would succeed, then would it not be fairer to allow him to participate in a new lottery, where he has a greater chance of getting the lesser punishment?

Perhaps there is a response to my examples on Lewis's behalf. In the first case, perhaps, you and I are actually each punished for planning a murder, not for firing bullets into bushes, where planning a murder always carries the chance that the intended victim does not in fact exist. Then, we each take the same chance of facing the penalty for murder; it just happens that your target exists while mine does not. In the second case, perhaps, your act is not really one of firing bullets into a person's flimsy home, but rather one of firing bullets into a randomly selected bush in some desert or other,

where it is very unlikely that a given desert will contain a person behind every bush. Then your act, unlike mine, actually carries a very small chance, commensurate with your reduced level of heartedness, of bringing you the penalty for murder.

I see two problems with this line of response. First, it necessitates a big change in our views about what kinds of acts receive punishment. We end up punishing acts distant from danger, like the act of making a plan, or overly abstract acts, like the act of shooting into some or other bush in some or other desert. Second, it smacks of confabulation. We start with the outcome of the penal lottery, then we go back and find a description of the circumstances of the lottery that makes the outcome seem more likely or less likely, depending upon the criminal's level of heartedness.

35.5 Obligations to the Distant Poor

Lewis wrote a short review of Peter Unger's book *Living High and Letting Die*.[7] Unger argues that our obligations to help impoverished people in other countries are much weightier than we ordinarily think. According to Unger, we are by nature inclined to think that we have greater obligations to people with whom we share "situations," or, equivalently, that we have lesser obligations to people from whom we are "separated." You might share a situation with a child drowning right in front of you, but be separated from a child starving on the other side of the world, and hence feel a strong obligation to save the nearby child, but no strong obligation to save the distant child. Unger thinks that the thesis about the psychological salience of situations and separation gives us reason to doubt our ordinary moral judgments, and to accept his radical conclusions in their place.

Lewis asks why we should not instead think the thesis to identify a principle which our ordinary moral judgments encode. Perhaps your sharing a situation with the nearby child explains why you really do have a strong obligation to save him from drowning, and perhaps your separation from the distant child explains why you really do not have a strong obligation to save her from starving. Then, says Lewis, our ordinary moral views are defended, and Unger's radical conclusions can be resisted (155–8).

Two things are worth noting. First, Lewis does not offer any deeper analysis of what it means for two people to share a situation, or to be separated. These are slippery and imprecise notions as they stand, but that does not prevent Lewis from taking them to do important explanatory work. Second, Lewis does not ask whether the notions have a good independent claim to moral significance; part of Unger's idea, I take it, is that whether someone is separated from or shares a situation with you just doesn't seem like it should make a big difference in your moral obligations. For Lewis, it seems, if the notions are helpful in unifying some of our ordinary moral judgments, then that is a strong reason to think that they bear moral significance indeed.

35.6 A Lewisian View in Social and Political Philosophy?

In his work in social and political philosophy, Lewis never presents a general theory, and never aligns himself with any theoretical perspective. To come up with a characterization of his overall views, we need to look for clues and commonalities in his treatments of his various narrower topics, and we need to speculate.

Each of Lewis's papers in social and political philosophy begins with a puzzle: a puzzle about what we really think (about penal substitution, for example), or about how one of our practices (like tolerating false opinions) can be justified, or about how we can achieve a shared goal (like

deterring a nuclear attack). Almost always, the constructive work in his papers is put towards resolving the puzzle in favor of our accepted opinions or practices. Toleration can be justified; it is not unjust to give harsher punishments for successful criminal attempts; we are not radically mistaken about our obligations to the distant poor; the paradox of deterrence can be resolved without any change to our initial evaluations of actions and intentions. "Do We Believe in Penal Substitution?" is the one paper that finds Lewis failing to vindicate a prevailing opinion or practice, and, even here, the conclusion is not that our judgments about penal substitution need to change, but rather that we need to look harder for their justification. "Buy like a MADman" is the one paper in which Lewis seriously considers a selection of different policies and recommends one over the others, and, even here, the recommendation takes the form of a sensible compromise between two opposing proposals. We can say with some certainty, then, that Lewis's approach to social and political philosophy is conservative. He pays a great deal of respect to the opinions of the sensible, earnest, well-intentioned ordinary person – one of "us" – and he is reluctant to question an accepted social practice.

(To call Lewis "conservative" in this sense is not to suggest that he is a political conservative. He gives little away concerning his political orientation, but he appears to have the temperament of a classical liberal. He has a strong interest in freedom of speech and expression, he always takes social policies to be justified by reference to the interests and entitlements of individual humans, and he is always more worried about how things could go wrong than he is excited about how they could be made more perfect.[8])

When evaluating ordinary opinions and practices, Lewis draws upon several styles of moral argument, in ways that make it impossible to place him within any recognizable school of moral theory. He disavows consequentialism and utilitarianism, but in his work on toleration and deterrence he evaluates acts and policies by reference to overall consequences.[9] He appeals to tacit treaties in his work on toleration, but contractarian strategies do not arise elsewhere. When speaking of toleration, he justifies an important civil right by appealing to rankings of social outcomes, but when speaking of punishment, he evaluates social policies by asking whether they respect prior truths about what individuals deserve. He can sound like a virtue ethicist when he criticizes utilitarianism, but he never argues from claims about virtue to claims about right acts or policies; the view in "Devil's Bargains," as we have seen, seems positively to rule out that kind of argument.[10]

We could take Lewis's treatments of narrower topics as installments in a progression towards a larger theory. Perhaps he would have liked to find a theoretical perspective, old or new, into which his various views in social and political philosophy could all comfortably fit. There is better reason, however, to see the variation within Lewis's styles of argument as evidence that he does not think that unified theories in social and political philosophy, or in moral philosophy generally, are needed, or even appropriate.[11] A revealing passage comes during a digression in "Devil's Bargains." Lewis is considering those who would retaliate with massive force to an attack upon their country, just out of a thirst for revenge.

> I myself would not despise them for being vengeful, though I think many moralists would. For I think their vengefulness is part of a package deal. It is inseparable from their love of their country and their solidarity with their countrymen. Conceptually inseparable, I am inclined to think…Be that as it may, surely the vengefulness and the solidarity are at least psychologically inseparable for people anything like ourselves. True, the vengeful fall short of being utilitarian saints. They are not motivated entirely by impersonal benevolence. But, as philosophers increasingly perceive, the utilitarian saint himself is a repellent figure. If it is the business of moral philosophy to sing his praises, moral philosophy only makes itself repellent. We should be less alienated from the things that real people really treasure. And these include the loyalties and affections from which vengefulness is inseparable. (206)

The passage suggests that our values and passions can be complicated and impure, but also that good moral philosophy engages sympathetically with those values and passions, rather than sitting distantly in judgment. It is easy to see how someone who accepts this picture of moral philosophy should doubt that all moral truths could be made to rest on a unified theoretical foundation; moral philosophy is answerable to human psychology, and human psychology is not that elegant.

If we can fairly attribute this picture to Lewis, then we can use it to bring together some further aspects of his work. Doing social and political philosophy well, the picture would suggest, is largely a matter of muddling through, making progress using any tools available; a theoretical approach that works for one problem might not work for another. Our first goal should be to support commonsensical opinions and accepted practices, on the picture, because when we try to overturn them, we run the risk of distancing ourselves from the things people really care about. It should not be surprising, on the picture, if we sometimes find good ethical reasons to cultivate in ourselves various kinds of psychological complexity, including a lack of self-transparency; with our complicated minds and messy values, we cannot presume that a clear focus on our goals is what will best help us to achieve them.[12]

I have ended with a putatively Lewisian picture of moral philosophy. Let me stress, again, that the picture is far from being a moral theory, and that the attempt to attribute it to Lewis is speculative. Lewis's work in social and political philosophy displays his familiar cleverness, originality, and idiosyncrasy, and it contains distinctive and underappreciated contributions on several important questions. But in this field, for once, Lewis is not a purveyor of theory.

Acknowledgments

I am very grateful for helpful comments from Ramon Das and Daniel Nolan, and for discussions with Catherine Bendeich, Dominic Keller, and David Ruschena.

Notes

1 The papers all appear in Lewis 2000, and all my page references are to this volume. There is also a further paper on deterrence (Lewis 1989); Lewis 1986 is an abridged version of this paper, and is the one that appears in *Papers in Ethics and Social Philosophy* – and the one to which I will refer here.

2 "Mill and Milquetoast" (published in 1989) is reprinted in a widely read collection on Mill (Dworkin 1997). For a recent discussion that refers to the paper in some detail, see Jacobson 2000, 280–1, 284. In making this judgment of Lewis's work in social and political philosophy, I do not intend to include his work on convention, which may well be counted a contribution to social philosophy, and is enormously influential.

3 The most extensive discussions of Lewis's papers are in Nolan 2005, chapter 8. There is also some good discussion in Weatherson 2010.

4 See, for example, Hursthouse 1999, chapter 1.

5 The argument of "Buy like a MADman" picks up on a discussion in "Devil's Bargains" (Lewis 2000, 210–16).

6 I borrow this illustration from Nolan 2005, 195.

7 Unger 1995.

8 See the liberalism articulated by Isaiah Berlin in Berlin (1969), especially pp. 167–72.

9 He says that "consequentialism is all wrong as everyday ethics, right as a limiting case" ("Devil's Bargains, Lewis 2000, 214).

10 For Lewis's criticisms of utilitarianism, see the passage quoted below, from p. 206, and see his comments in 1986, pp. 127–8.

11 See Nolan 2005, 189.

12 See my earlier comments about the implications for moral psychology of the argument of "Buy like MADman"; and see Lewis's comments about the psychology of toleration in "Mill and Milquetoast" (185).

References

Berlin, I. (1969). *Four essays on liberty*. Oxford: Oxford University Press.

Dworkin, G. (Ed.) (1997). *Mill's* On liberty*: critical essays*. Lanham, MD: Rowan and Littlefield.

Jacobson, D. (2000). Mill on liberty, speech, and the free society. *Philosophy and Public Affairs*, 29(3), 276–309.

Hursthouse, R. (1999). *On virtue ethics*. Oxford: Oxford University Press.

Lewis, D. (1986). *On the plurality of worlds*. Oxford: Blackwell.

Lewis, D. (1989). Finite counterforce. In H. Shue (Ed.), *Nuclear deterrence and moral restraint*, pp. 51–114. Cambridge: Cambridge University Press.

Lewis, D. (2000). *Papers in ethics and social philosophy*. Cambridge: Cambridge University Press.

Nolan, D. (2005). *David Lewis*. Montreal and Kingston: McGill-Queen's University Press.

Unger, P. (1995). *Living high and letting die: our illusion of innocence*. Oxford: Oxford University Press.

Weatherson, B. (2010). David Lewis. In E.N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2010 edition), http://plato.stanford.edu/archives/sum2010/entries/david-lewis/, accessed June 17, 2014.

Bibliography of the Work of David Lewis

Compiled by Stephanie R. Lewis, July 7, 2012

1966

An argument for the identity theory. *Journal of Philosophy*, 63, 17–25.
Percepts and color mosaics in visual experience. *Philosophical Review*, 75, 357–68.
Scriven on human unpredictability. *Philosophical Studies*, 17, 69–74 (J.S. Richardson, co-author).

1968

Counterpart theory and quantified modal logic. *Journal of Philosophy*, 65, 113–26.
Finitude and infinitude in the atomic calculus of individuals. *Noûs*, 2, 405–10 (W. Hodges, co-author).

1969

Policing the Aufbau. *Philosophical Studies*, 20, 13–17.
Review of Capitan & Merrill (Eds.), *Art, mind, and religion*. *Journal of Philosophy*, 66, 22–7.
Lucas against mechanism. *Philosophy*, 44, 231–33.
Convention: a philosophical study. Cambridge, MA: Harvard University Press, second printing 1974.

1970

Anselm and actuality, *Noûs*, 4, 175–88.
How to define theoretical terms. *Journal of Philosophy*, 67, 427–46.

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

Holes. *Australasian Journal of Philosophy*, 48, 206–12 (S.R. Lewis, co-author).
Nominalistic set theory. *Noûs*, 4, 225–40.
General semantics. *Synthese*, 22, 18–67.

1971

Immodest inductive methods. *Philosophy of Science*, 38, 54–63.
Counterparts of persons and their bodies. *Journal of Philosophy*, 68, 203–11.
Analog and digital. *Noûs*, 5, 321–7.
Completeness and decidability of three logics of counterfactual conditionals. *Theoria*, 37, 74–85.

1972

Utilitarianism and truthfulness. *Australasian Journal of Philosophy*, 50, 17–9.
Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, 50, 249–58.

1973

Counterfactuals. Oxford: Basil Blackwell.
Causation. *Journal of Philosophy*, 70, 556–7.
Counterfactuals and comparative possibility. *Journal of Philosophical Logic*, 2, 418–46.
Lingue e lingua, *Versus*, 4, 2–21.

1974

Semantic analyses for dyadic deontic logic. In S. Stenlund (Ed.), *Logical theory and semantic analysis: essays dedicated to Stig Kanger on his fiftieth birthday*. Dordrecht: Reidel.
Spielman and Lewis on inductive immodesty. *Philosophy of Science*, 41, 84–5.
Intensional logics without iterative axioms. *Journal of Philosophical Logic*, 3, 457–66.
Radical interpretation. *Synthese*, 23, 331–44.
'Tensions. In M.K. Munitz & P.K. Unger (Eds.), *Semantics and philosophy*. New York: New York University Press.

1975

Languages and language. In K. Gunderson (Ed.), *Minnesota studies in the philosophy of science*, Vol. VII. Minneapolis: University of Minnesota Press.
Adverbs of quantification. In E.L. Keenan (Ed.), *Formal semantics of natural language*. Cambridge: Cambridge University Press.
Review of Olson and Paul, *Contemporary philosophy in Scandinavia*. *Theoria*, 41, 39–60 (S.R. Lewis, co-author).

1976

Convention: reply to Jamieson. *Canadian Journal of Philosophy*, 6, 113–20.
The paradoxes of time travel. *American Philosophical Quarterly*, 13, 145–52.

Probabilities of conditionals and conditional probabilities. *Philosophical Review*, 85, 297–315.
Survival and identity. In A.O. Rorty (Ed.), *The identities of persons*. Berkeley: University of California Press.

1977

Possible-world semantics for counterfactual logics: a rejoinder. *Journal of Philosophical Logic*, 6, 359–63.

1978

Truth in fiction, *American Philosophical Quarterly*, 15, 37–46.
Reply to McMichael, *Analysis*, 38, 85–6.

1979

A problem about permission. In E. Saarinen *et al*. (Eds.), *Essays in honour of Jaakko Hintikka*. Dordrecht: Reidel.
Prisoners' Dilemma is a Newcomb Problem. *Philosophy and Public Affairs*, 8, 235–40.
Counterfactual dependence and time's arrow. *Noûs*, 13, 455–76.
Scorekeeping in a language game. *Journal of Philosophical Logic*, 8, 339–59.
Attitudes *de dicto* and *de se*. *Philosophical Review*, 88, 513–43.
Lucas against mechanism II. *Canadian Journal of Philosophy*, 9, 373–6.

1980

A subjectivist's guide to objective chance. In R.C. Jeffrey (Ed.), *Studies in inductive logic and probability*, Vol. II. Berkeley: University of California Press.
Mad pain and Martian pain. In N. Block, D.T. Langendoen & J.J. Katz (Eds.), *Readings in philosophy of psychology*, Vol. I. Cambridge, MA: Harvard University Press.
Index, context, and content. In S. Kanger & S. Øhman (Eds.), *Philosophy and grammar*. Dordrecht: Reidel.
Veridical hallucination and prosthetic vision. *Australasian Journal of Philosophy*, 58, 239–49.

1981

Causal decision theory, *Australasian Journal of Philosophy*, 59, 5–30.
Ordering semantics and premise semantics for counterfactuals, *Journal of Philosophical Logic*, 10, 217–34.
What puzzling Pierre does not believe. *Australasian Journal of Philosophy*, 59, 283–9.
Why ain'cha rich?, *Noûs*, 15, 377–80.
Are we free to break the laws?, *Theoria*, 47, 113–21.

1982

Logic for equivocators. *Noûs*, 16, 431–41.
"Whether" Report. In T. Pauli *et al*. (Eds.), *320311: Philosophical essays dedicated to Lennart Åqvist on his fiftieth birthday*. Uppsala: Filosofiska Studier.
Censored vision (written under the name "Bruce LeCatt"). *Australasian Journal of Philosophy*, 60, 158–62.

1983

Philosophical papers, Vol. I. Oxford: Oxford University Press. Reprinted papers, with new postscripts and a new introduction.
The papers in this volume are:
Holes
Anselm and Actuality with new postscripts
Counterpart Theory and Quantified Modal Logic with new postscripts
Counterparts of Persons and Their Bodies
Survival and Identity with new postscripts
How to Define Theoretical Terms
An Argument for the Identity Theory with additions
Radical Interpretation with new postscripts
Mad Pain and Martian Pain with new postscripts
Attitudes *De Dicto* and *De Se* with new postscripts
Languages and Language
General Semantics with new postscripts
Scorekeeping in a Language Game
'Tensions
Truth in Fiction with new postscripts
Individuation by acquaintance and by stipulation. *Philosophical Review*, 92, 3–32.
Extrinsic properties. *Philosophical Studies*, 44, 197–200.
New work for a theory of universals. *Australasian Journal of Philosophy*, 61, 343–77.
Levi against U-maximization. *Journal of Philosophy*, 80, 531–4.

1984

Putnam's paradox, *Australasian Journal of Philosophy*, 62, 221–36.
Devil's bargains and the real world. In D. MacLean (Ed.), *The security gamble: deterrence in the nuclear age*. Totowa, NJ: Rowman and Allenheld.

1986

On the plurality of worlds. Oxford: Basil Blackwell, reissued by Blackwell, 2001.
Philosophical papers, Vol. II. Oxford: Oxford University Press. Reprinted papers with two hitherto unpublished papers, "Causal Explanation" and "Events," with new postscripts, a new

introduction, and a bibliography which includes abstracts of *Convention*, *Counterfactuals*, and *On the Plurality of Worlds*, and of all the papers published through 1986 which are not reprinted in *Philosophical Papers* I or II.

The papers in this volume are:

Counterfactuals and Comparative Possibility
Counterfactual Dependence and Time's Arrow with new postscripts
The Paradoxes of Time Travel
A Subjectivist's Guide to Objective Chance with new postscripts
Probabilities of Conditionals and Conditional Probabilities with new postscript
Causation with new postscripts
Causal Explanation first published here
Events first published here
Veridical Hallucination and Prosthetic Vision with new postscript
Are We Free To Break the Laws?
Prisoners' Dilemma is a Newcomb Problem
Causal Decision Theory with new postscripts
Utilitarianism and Truthfulness

Against structural universals. *Australasian Journal of Philosophy*, 64, 25–46.
Buy like a MADman, use like a NUT. *QQ*, 6, 5–8.
Comment on Armstrong and Forrest. *Australasian Journal of Philosophy*, 64, 92–3.
Probabilities of conditionals and conditional probabilities II. *Philosophical Review*, 95, 581–9.
Causal explanation. In D. Lewis, *Philosophical papers*, Vol. II. Oxford: Oxford University Press.
Events. In D. Lewis, *Philosophical papers*, Vol. II. Oxford: Oxford University Press.

1987

The punishment that leaves something to chance, *Proceedings of the Russellian Society* (University of Sydney), 12, 81–97.

1988

Ayer's first empiricist criterion of meaning: why does it fail? *Analysis*, 48, 1–3.
Rearrangement of particles: reply to Lowe. *Analysis*, 48, 65–72.
Statements partly about observation. *Philosophical Papers*, 17, 1–31.
The trap's dilemma. *Australasian Journal of Philosophy*, 66, 220–3.
Desire as belief, *Mind*, 97, 323–32.
Relevant implication, *Theoria*, 54, 162–74.
Vague identity: Evans misunderstood. *Analysis*, 48, 128–30.
What experience teaches, *Proceedings of the Russellian Society* (University of Sydney), 13, 29–57. Reprinted in D. Lewis, *Papers in metaphysics and epistemology*. Cambridge: Cambridge University Press.

1989

Die Identität von Körper und Geist. Frankfurt: Klostermann. (An Argument for the Identity Theory, Psychophysical and Theoretical Identifications, and Mad Pain And Martian Pain, trans. and with an afterword by A. Kemmerling.)

Finite counterforce. In H. Shue (Ed.), *Nuclear deterrence and moral restraint*. Cambridge: Cambridge University Press.
Academic appointments: why ignore the advantage of being right? In *Ormond Papers* (Ormond College, University of Melbourne, 1989). Reprinted in D. Lewis, *Papers in ethics and social philosophy*. Cambridge: Cambridge University Press.
Mill and milquetoast. *Australasian Journal of Philosophy*, 67, 152–71.
Dispositional theories of value. *Proceedings of the Aristotelian Society*, Suppl. Vol. 63, 113–37.
Review of John Bigelow, *The reality of numbers*. *Australasian Journal of Philosophy*, 67, 487–9.

1990

Noneism or allism? *Mind*, 99, 23–31.

1991

Parts of classes. Oxford: Basil Blackwell.

1992

Meaning without use: reply to Hawthorne. *Australasian Journal of Philosophy*, 70, 106–10.
Critical notice of Armstrong, *A combinatorial theory of possibility*. *Australasian Journal of Philosophy*, 70, 211–24.

1993

Counterpart theory, quantified modal logic, and extra argument places. *Analysis*, 53, 69–71.
Many, but almost one. In K. Campbell, J. Bacon & L. Reinhardt (Eds.), *Ontology, causality and mind: essays on the philosophy of D.M. Armstrong*. Cambridge: Cambridge University Press.
Mathematics is megethology. *Philosophia Mathematica*, 3, 3–23.
Evil for freedom's sake?, *Philosophical Papers*, 22, 149–72.

1994

Reduction of mind. In S. Guttenplan (Ed.), *A companion to philosophy of mind*. Oxford: Blackwell.
Humean supervenience debugged. *Mind*, 103, 473–90.

1995

Should a materialist believe in qualia? *Australasian Journal of Philosophy*, 73, 140–4.
Ern Malley's namesake. *Quadrant*, 39, 14–5.

1996

Desire as belief II, *Mind*, 105, 303–13.
Review of R. Casati and A. Varzi, *Holes*. *Philosophical Review*, 105, 77–9 (S.R. Lewis, co-author).
Elusive knowledge. *Australasian Journal of Philosophy*, 74, 549–67.
Maudlin and modal mystery. *Australasian Journal of Philosophy*, 74, 683–4.
Illusory innocence? (review of P. Unger, *Living high and letting die*). *Eureka Street*, 6 (10) December, 35–6.

1997

Finkish dispositions. *Philosophical Quarterly*, 47, 143–58.
Naming the colours. *Australasian Journal of Philosophy*, 75, 325–42.
Do we believe in penal substitution? *Philosophical Papers*, 26, 203–9.

1998

Papers in philosophical logic. Cambridge: Cambridge University Press.
The papers in this volume are:
- Adverbs of Quantification
- Index, Context and Content
- 'Whether' Report
- Probabilities of Conditionals and Conditional Probabilities II
- Intensional Logics without Iterative Axioms
- Ordering Semantics and Premise Semantics for Counterfactuals
- Logic for Equivocators
- Relevant Implication
- Statements Partly About Observation
- Ayer's First Empiricist Criterion of Meaning: Why does It Fail?
- Analog and Digital
- Lucas against Mechanism
- Lucas against Mechanism II
- Policing the *Aufbau*
- Finitude and Infinitude in the Atomic Calculus of Individuals (with W. Hodges)
- Nominalistic Set Theory
- Mathematics Is Megethology

A world of truthmakers? (review of D.M. Armstrong, *A world of states of affairs*), published under an incorrect title in *Times Literary Supplement*, 4950 (February 13), 30.
Defining "Intrinsic." *Philosophy and Phenomenological Research*, 58, 333–45 (R. Langton, co-author).
Void and object. In J. Collins, N. Hall & L.A. Paul (Eds.), *Causation and counterfactuals*. Cambridge, MA: MIT Press.

1999

Papers in metaphysics and epistemology. Cambridge: Cambridge University Press.
The papers in this volume are:
- New Work for a Theory of Universals
- Putnam's Paradox
- Against Structural Universals
- A Comment on Armstrong and Forrest (AJP publication is titled Comment on Armstrong and Forrest)
- Extrinsic Properties
- Defining 'Intrinsic' (with R. Langton)
- Finkish Dispositions
- Noneism or Allism?
- Many, But Almost One
- Casati and Varzi on Holes (with S.R. Lewis)
- Rearrangement of Particles: Reply to Lowe
- Armstrong on Combinatorial Possibility
- A World of Truthmakers?
- Maudlin and Modal Mystery
- Humean Supervenience Debugged
- Psychophysical and Theoretical Observations
- What Experience Teaches
- Reduction of Mind
- Should a Materialist Believe in Qualia?
- Naming the Colours
- Percepts and Color Mosaics in Visual Experience
- Individuation by Acquaintance and by Stipulation
- Why Conditionalize? previously unpublished
- What Puzzling Pierre Does Not Believe
- Elusive Knowledge

Zimmerman and the spinning sphere. *Australasian Journal of Philosophy*, 77, 209–12.
Why conditionalize? Written in 1972 as a course handout, introduction written in 1997, and printed for the first time in D. Lewis, *Papers in metaphysics and epistemology*. Cambridge: Cambridge University Press.

2000

Papers in ethics and social philosophy. Cambridge: Cambridge University Press.
The papers in this volume are:
- Semantic Analysis for Dyadic Deontic Logic
- A Problem about Permission
- Reply to McMichael
- Why Ain'cha Rich?
- Desire as Belief
- Desire as Belief II

Dispositional Theories of Value
The Trap's Dilemma
Evil for Freedom's Sake?
Do We Believe in Penal Substitution?
Convention: Reply to Jamieson
Meaning without Use: reply to Hawthorne
Illusory Innocence?
Mill and Milquetoast
Academic Appointments: Why Ignore the Advantage of Being Right?
Devil's Bargains and the Real World
Buy Like a MADman, Use Like a NUT
The Punishment That Leaves Something to Chance
Scriven on Human Unpredictability (with J.S. Richardson)

Causation as influence (abridged version). *Journal of Philosophy*, 97, 182–97.

2001

Sleeping Beauty: reply to Elga, *Analysis*, 61, 171–6.
Forget about the "correspondence theory of truth." *Analysis*, 61, 275–80.
Marshall and Parsons on "intrinsic." *Philosophy and Phenomenological Research*, 63, 353–5 (R. Langton, co-author).
Redefining "intrinsic." *Philosophy and Phenomenological Research*, 63, 381–98.
Truthmaking and difference-making. *Noûs*, 35(4), 601–15.

2002

Tharp's third theorem. *Analysis*, 62, 95–7.
Tensing the copula. *Mind*, 111, 1–13.

2003

Things qua truthmakers. In H. Lillehammer & G. Rodriguez-Pereyra (Eds.), *Real metaphysics: essays in honour of D.H. Mellor*. London: Routledge, pp. 25–38.
Postscript to "things qua truthmakers": negative existentials (G. Rosen, co-author). In H. Lillehammer & G. Rodriguez-Pereyra (Eds.), *Real metaphysics: essays in honour of D.H. Mellor*. London: Routledge, pp. 39–42.
Rights to rights. *Theoria*, 69, 160–5.

2004

Tensed quantifiers. In D. Zimmerman (Ed.), *Oxford studies in metaphysics*, Vol. 1. Oxford: Oxford University Press, pp. 3–14.
How many lives has Schrödinger's cat? *Australasian Journal of Philosophy*, 82, 3–22.

Causation as influence (unabridged version). In J. Collins, N. Hall & L.A. Paul (Eds.), *Causation and counterfactuals*. Cambridge, MA: MIT Press.

Void and object. In J. Collins, N. Hall & L.A. Paul (Eds.), *Causation and counterfactuals*. Cambridge, MA: MIT Press.

Letters to Priest and Beall. In G. Priest, J.C. Beall & B. Armour-Garb (Eds.), *The law of non-contradiction*. Oxford: Oxford University Press, pp. 176–7.

2007

Materialismus und BewusstseinI. Frankfurt: Klostermann/Seminar. (What Experience Teaches, Reduction of Mind, and Should a Materialist Believe in Qualia? trans. U. Haas-Spohn.)

Quasi-realism is fictionalism, In M.E. Kalderon (Ed.), *Moral fictionalism*. Oxford: Oxford University Press.

Divine evil. In L.M. Antony (Ed.), *Philosophers without gods: meditations on atheism and the secular life*. Oxford: Oxford University Press.

2009

Ramseyan humility. In D. Braddon-Mitchell & R. Nola (Eds.), *Conceptual analysis and philosophical naturalism*. Cambridge, MA: MIT Press.

2015

States of affairs reconsidered. In B. Loewer & J. Schaffer (Eds.), *A Companion to David Lewis*. Oxford: Wiley-Blackwell, Chapter 2.

Reply to Dana Scott. In B. Loewer & J. Schaffer (Eds.), *A Companion to David Lewis*. Oxford: Wiley-Blackwell, Chapter 3.

Index

Abbott, Barbara, 347–348
absence(s)
  events *see* causation, by omission
  holes *see* hole(s), as absences
accommodation *see* presupposition, accommodation
action, 434, 534–554
actuality (actual world), 19, 89
Adams, Robert, 19, 208
Adams's Thesis, 427–428, 432
adjectives, gradable, 482–483
adverbs of quantification, 337, 352–353
Albert, David, 192–193, 285–286
Albert the Great, 63
Albert of Saxony, 67, 68
Al-Ghazali, 70
Allen, James, 354, 358
Alter, Torin, 526
analyticity, 34, 144, 504–508, 546–547
anaphora, 350–351
  donkey, 331, 337
Anderbois, Scott, 357
anti-haecceitism, 101–102
anti-realism, 62–63
  see also realism
Appelt, Douglas, 358
a priori, 40–41, 55, 87–89
Aquinas, Thomas, 63, 73–74
Aristotle, 62–64, 67, 69, 72–73, 221
Arló Costa, Horacio, 454
Armstrong, David, 141, 147–148, 159, 164–165, 234, 251–252, 389–390, 512
Arnauld, Antoine, 67
Arntzenius, Frank, 135n
Asher, Nicholas, 362
assertability, 427, 432
asymmetry
  of context-shift, 484
    see also context
  counterfactual, 142–143, 298
    see also counterfactual(s)
  see also symmetry
attitudes, propositional, 332, 335
  embedding of, 332
  see also belief; *de re* attitudes; *de se* attitudes; desire
Augustine, 67
Austin, John, 475n
Avicenna, 67
axiom(s)
  of choice, 465–467
  of mereology, 253–254
    see also composition; mereology
  of rational choice *see* decision theory
  of set theory, 468–469

Bacon, Francis, 71
Baker, Lynne Rudder, 251
Baker, Mark, 323
Baldridge, Jason, 336
Barnes, Elizabeth, 253
Barwise, Jon, 88
Bäuerle, Richard, 340
Baxter, Donald, 258, 463
Bayesianism, 30–31, 36, 38, 512
  conditionalization, update by, 30
  imaging, 448
  Jeffrey conditionalization, 426, 429, 436, 438–440
  logical omniscience, 36, 514
  prior credence(s), 30–31

A Companion to David Lewis, First Edition. Edited by Barry Loewer and Jonathan Schaffer.

© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

Beaver, David, 349–350, 353, 357
Beebee, Helen, 234–235, 272–274
behaviorism, 504–505, 538
belief, 198, 511–515
  content of, 195–198
    see also de re attitudes; *de se* attitudes; semantic theories
  degree of (credence) *see* Bayesianism
  de re see de re attitudes
  de se see de se attitudes
  and desire *see* decision theory; Desire-as-Belief Thesis; functionalism
  initial *see* Lewis, David, methodology of, conservatism
  and value, 541–542
Benacerraf, Paul, 462
Bennett, Jonathan, 85
Bennett, Karen, 251, 258, 260
Bergstrom, Karl, 319
Best System Account (BSA)
  of chance, 189–192, 284
  and explanatory power, 271
  indexicality, 118–119, 121–122
  language of, 118, 121–126
  of laws of nature, 117–130, 141–142, 189–192, 262–276
  metrics for, 118–119, 121–125, 192–193, 267–269
    see also simplicity
  objectivity, 118–119, 121
  pragmatism, 192–193, 267–269
  problem of initial conditions, 269–270, 272–273
  relativity, 118–119, 124–125
  rigidity, 118–119, 123–124
  triviality, 195
  two guiding ideas, 265–267
Bigelow, John, 101, 103, 111, 284–285
binding, 337
Bird, Alexander, 141, 497
Blackburn, Simon, 514
Blome-Tillmann, Michael, 478, 483
Boër, Steven, 350
Boethius, 72
Boolos, George, 459–460, 467–469
Boyle, Robert, 63, 69–70
Braddon-Mitchell, David, 38, 509
brain state(s), 508–509, 512–514
Bricker, Phillip, 183n, 184n, 185n
Burge, Tyler, 323, 515
Burgess, John, 383
Buridan, John, 66–67
Büring, Daniel, 352
Butterfield, Jeremy, 44, 113–114

Cai, Fang, 319
Callender, Craig, 119–120
Cambridge change, 103
Cameron, Ross, 258
Canberra Plan, 32–33, 40, 505–508
  see also functionalism
Cantor, Georg, 461–462
Carnap, Rudolf, 34, 80, 83, 505–508
Casati, Roberto, 42–43, 50–54
Castaneda, Hector-Neri, 89–90
Cath, Yuri, 526–527
causation, 68–71, 85, 492–494
  agent, 69–70
  as ancestral of causal dependence, 298–300
  as ancestral of quasi-dependence, 301
  causal dependence, 298–300, 306–307
  double prevention, 302
  indeterministic, 300–301
  as influence, 303–305
  as intrinsic relation, 107–108, 301
  kinds of, 68–70
  occasionalism, 70–71
  by omission, 301
  preemption, 300–301, 304
  by prevention, 301
  probabilistic, 281–282, 288–289
  transitivity of, 300
  trumping, 302–303
  see also laws of nature
centered world(s) *see de se* attitudes, and centered worlds
C-fiber, 85, 87, 504, 508–512
  see also pain
Chalmers, David, 88–89, 128, 373, 403, 515–516, 525
chance, 107, 193, 278–286
  and determinism, 278–279, 286–292
  and frequency, 190
  as guide to credence *see* Principal Principle
  indexing, 280–281
  as magnitude, 281–282, 288–289
  across time, 280–281
  see also Best System Account; Principal Principle
character, 337–338
Chen, Shenglu, 319
Chierchia, Gennaro, 332, 338
Chomsky, Noam, 322–323
Churchland, Paul, 108
Clark, Robin, 316
closeness of worlds *see* similarity, between worlds
coexistence, 498–502
Cohen, Jonathan, 119–120
Cohen, Stewart, 478
collective predicate *see* plurality
Collins, John, 111, 284–285, 454
color, 27–29, 34, 63, 87
  see also knowledge argument
common ground, 357–358, 361
communication
  as coordination problem, 318–319
compartmentalization, 486
composition
  as identity, 172, 256–257
  non-mereological, 171, 228, 252
    see also states of affairs, Lewis's critique of
  ontological innocence, 256–258
  uniqueness, 170, 251–252
  unmysterious, 258–259
  unrestricted, 163, 252–256
  see also mereology

compositionality, 322, 333
concept(s), 514
conceptual analysis, 41–42
conditional excluded middle, 84, 427
constitution, 173–174, 227–228, 251
construction, 15–16, 382–384
  see also composition
content *see* belief, content of; meaning
  at-issue *see* presupposition; question under discussion (QUD)
context, 92–93, 338, 357–358
  see also knowledge, contextualism
convention, 91–92, 319–326, 368
  origin of, 320–321, 325–326
  and routine, 324
  of truthfulness and trust, 334
  see also meaning
coordination problem, 315–319
  focal point, 320
  see also game theory
correspondence theory of truth *see* truth, correspondence theory
Cotnoir, Aaron, 258
counterfactual(s), 84–85, 142–143, 210–212, 412–424
  asymmetry *see* asymmetry, counterfactual
  backtracking, 298–300, 309
    see also asymmetry, counterfactual
  counterfactual dependence, 298
  Goodman's theory, 412–415
  non-backtracking, 297–298
counterpart relation, 173, 222–225
  modal vs counterpart relation temporal, 244–248
  multiplicity, 224–227
  and similarity, 222–224
  vs transworld identity, 122
  see also counterpart theory
counterpart theory, 72, 81
  and context, 225–226
  and essentialism, 15–17, 174, 222–224
  Kripke's Humphrey objection to, 81
  and semantics, 335
  and truthmaking, 173–180, 227–230
Coventry, Angela, 234–235
creation, 19–20
credence *see* Bayesianism
Cresswell, Max, 338
Curley, Edward, 212
Curry, Haskell, 335

Darwall, Stephen, 540
Davidson, Donald, 389
decision, 197
decision theory, 433–441, 446–458
  causal, 305–307, 427–428, 447–451
  dependency hypotheses, 305–306, 450–452
  evidential, 447
  see also Desire-as-Belief Thesis; game theory
Dedekind, Richard, 462–463
dependence
  causal *see* causation, causal dependence
  context *see* context
  counterfactual *see* counterfactual(s), counterfactual dependence
  metaphysical *see* fundamentality; grounding; supervenience; truthmaker(s)
de re attitudes, 91, 405–406
DeRose, Keith, 207, 478, 483, 486
Descartes, René, 63–64, 69, 71, 73
de se attitudes, 89–91, 515
  and centered worlds, 400, 403–405
  and controlled infinitives, 338
  irreducible to *de dicto*, 401–402
  self-ascription, 90–91, 400, 403, 406–409
desire
  and belief *see* decision theory; Desire-as-Belief Thesis; functionalism
  and value, 536–546
Desire-as-Belief Thesis, 433–442, 452–458
  see also decision theory
determiner, 352, 482–484
  see also domain restriction; quantifier(s)
determinism, 271, 278–279
deterrence, 553–555
DeVault, David, 347–348
Devitt, Michael, 323
diagonalization, 20
  see also Cantor, Georg; set theory
Diodorus, 72
dispositions, 69, 234, 307–308, 415
Divers, John, 151
dogmatism, epistemic, 104
domain restriction, 352–353
donkey sentences *see* anaphora, donkey
Dorr, Cian, 185n
Dowty, David, 334
Dretske, Fred, 475
dualism, 419, 520, 530
  see also materialism
duplication, 103, 106, 113, 127, 135n, 145–146, 186n, 374, 521
  see also properties, intrinsic

Earman, John, 45, 270
Edgington, Dorothy, 85, 288–289, 428
Egli, Urs, 340
Elga, Adam, 474
Ellis, Brian, 497
empiricism, 64, 138–141
epistemology *see* knowledge
essentialism, 221–223
  essentiality of origin, 147
  nomic *see* laws of nature, necessitarianism about
ethics *see* value; virtue ethics
events, 297
  alterations, 303
  distinct, 297, 299, 389
  fragile, 297
  omissions, 31
evidence, 28, 30–32, 36–38, 282–283, 290, 429, 474–477, 479–482, 528, 542
  admissible, 282–283
  see also knowledge; Principal Principle

evil, divine, 212–213
evil, problem of *see* theodicy
expectation, social, 316
  and knowledge of meaning, 321
experience, 522–526
explanation, 26, 62–63, 309, 412–413
  explanatory power, 18–19
    see also Best System Account, and explanatory power; grounding; truthmaker(s)
  non-causal, 26

Field, Hartry, 19, 367–370, 376
Fine, Kit, 155
finks *see* dispositions
focus, 351–352, 361–362
Fodor, Jerry, 331–332
foreknowledge, divine, 209–212
forms, theory of *see* hylomorphism
Forrest, Peter, 159, 216
four-dimensionalism *see* persistence, perdurantism
freedom, 209–212
Frege–Geach problem, 433
Frege, Gottlob, 19, 90, 334, 340, 400, 403, 406, 465, 467–468, 505
Frege's puzzle *see* belief, content of
French, Steven, 26
frequentism, 189
function, 336
functionalism, 86–87, 523
  analytic, 504–508
  psychofunctionalism, 507
  realizer, 509–511
  role, 509, 511
  see also Lewis, David, conceptual analysis, Ramseyfication; mathematics, structuralism about
fundamentality, 103, 160–161
  see also grounding; properties, natural; supervenience; truthmaker(s)

game theory, 317–318, 324
  Nash equilibrium, 317
Gauker, Christopher, 347–348
Gibbard, Allan, 173–174, 227–228, 306, 427–428
Gilbert, Margaret, 324
Ginzburg, Jonathan, 357
God, 7, 65–68, 70–74, 206–219
Gödel, Kurt, 466
Goodman, Nelson, 104, 412–415
Grafen, Alan, 325
grammar
  categorial, 335–336
  universal, 322–324
Grice, Paul, 8, 321, 328, 354, 358, 360, 484
Groenendijk, Jeroen, 352
grounding, 73–74, 103, 161, 181, 259, 374
  and counterfactual dependence, 121–122, 125–126
  see also truthmaker(s)

haecceitism *see* anti-haecceitism
Hájek, Alan, 431–432, 440
Hall, Ned, 103, 264, 267, 282–284, 301–302
Halliday, Michael, 362
Halpern, Joseph, 301–302
Halpin, John, 119–120
Ham, Fritz, 334
Hamblin, Charles, 352
Harris, Randy, 329
Harris, Zellig, 328
Harper, William, 306, 427–428
Hasan, Ruqaiya, 362
Haslanger, Sally, 103, 243, 255
Hawley, Katherine, 254, 256
Hawthorne, John, 156, 378, 475, 497
Hazen, Allen, 185n, 215, 383, 459–461, 463–469
Heim, Irene, 93, 331, 336, 337, 340, 347–348, 483
Hilbert, David, 467
Hintikka, Jaakko, 72, 483
Hitchcock, Christopher, 301–302
Hobbes, Thomas, 64
Hoefer, Carl, 45–46, 119–120, 126, 130, 280
hole(s), 42–58
  as absences, 49
  eliminativism about, 47–48, 49–52, 54, 58
  functionalism about, 52, 56–57
  as immaterial beings, 42–43, 50–51, 53
  as material beings, 43, 49, 51, 53
  nominalism about, 48, 50, 56
Holton, Richard, 93
Hookway, Christopher, 485
Humberstone, Lloyd, 113, 502n
Humean supervenience, 102–103, 140–141, 189–202, 418–420
  and classical physics, 112–114
  contingency of, 106, 233
  and *de re* modality, 232–235
  and modal combinatorialism, 105–106
  and nomological reductionism, 106–107, 234–235
  and persistence, 238–242
  and quantum physics, 108, 189
Hume, David, 71, 103–105, 139–141, 230, 234–235, 412–413, 493, 537
Hume's dictum (HD), 71, 138–148, 171–172, 228, 233–235, 264
hylomorphism, 64–67, 72–74, 76
hyperintensionality, 88–89, 91–92, 478–479

Ichikawa, Jonathan, 478, 483
identity, 227–228, 256
  across time *see* persistence
  across worlds *see* counterpart theory, and essentialism
  contingent, 173–174, 509–512
    see also counterpart theory
identity theory of mind, 508–509, 524
imaging, 448
implicature, conversational, 321–322, 355
indeterminacy
  metaphysical, 20–21, 373–374
  referential, 28, 369–370, 378
  see also semantic theories, eligibility

indeterminism, 212, 278, 288–289
index, 332, 338
  see also context
indexical(s), 482–483
Induction, New Riddle of, 103
infallibilism, 477
  see also knowledge
infinity, 72–73, 465–466, 468–469
influence, causal *see* causation, as influence
informativeness, 265, 270, 272–273
  see also Best System Account
initial conditions, 263
instrumentalism, scientific, 27
intension, 40–41, 332
intention *see* belief; desire; decision theory; functionalism
intentionality *see* meaning; semantics
internalism, judgment, 540
interpretation, 318
interpretationism, 367–368
  see also semantic theories
Ismael, Jenann, 285–286

Jackendoff, Ray, 332, 335
Jackson, Frank, 40–41, 88–89, 109, 389, 402, 507–511, 516, 525, 530
Jacobs, Roderick, 329
Jacobson, Daniel, 537
Jacobson, Pauline, 336
Jasinskaja, Katja, 357
Jeffrey, Richard, 30, 426–430, 434, 438
  see also Bayesianism, Jeffrey conditionalization
Johnston, Mark, 238
Joyce, James, 112, 514
justification *see* knowledge

Kahn, Herman, 10
Kamp, Hans, 11, 329, 331, 337, 340
Kant, Immanuel, 491–494, 498–499, 501–502
Kaplan, David, 19, 88–89, 338, 399, 482–483
Karttunen, Lauri, 347–349
Katz, Jerrold, 331–332
Kaufmann, Stefan, 433
Kehler, Andrew, 362
Keill, John, 63
Kenny, Antony, 209
Kim, Jaegwon, 511
kind, natural *see* properties, natural
Kment, Boris, 85
know-how *see* knowledge, how
knowledge, 474–486
  common, 320, 322
  contextualism, 475–478, 481–486, 528–529
  foreknowledge, divine *see* foreknowledge, divine
  how, 407, 519, 526–528
  monism about knowledge relations, 481–482
  relevant alternatives, 473–480, 483–485
knowledge argument, 402, 516, 524–525, 526–528
Korsgaard, Christine, 455
Koslicki, Kathrin, 260
Kratzer, Angelika, 147, 335, 336, 340
Kripke, Saul, 80–87, 89, 147, 208, 223, 335, 351, 359, 505, 526, 530

Lachmann, Michael, 319
Lakoff, George, 332–333, 335
lambda abstraction *see* binding
Landman, Fred, 336
Lange, Marc, 281
Langton, Rae, 374, 494
Langtry, Bruce, 214–215
language(s), 91–93, 333–334
  faculty, 322–323
Lascarides, Alex, 362
laws of nature, 262
  governing conception of, 268–269
  necessitarianism about, 497
  see also Best System Account
Leibniz, Gottfried, 72–74, 493–494
Lepore, Ernie, 324, 330
Levi, Isaac, 429, 454
Lewis, David
  conceptual analysis, 296–297
    conservatism, 29–32, 41, 231–232
    intuition, 19–20, 232, 296–297
    naturalism, 82–88
    Ramseyfication, 33–35, 83–84, 495–497, 504–508
    science, 26–27, 83–84
    systematicity, 25–26, 36
    theoretical virtues, 35–37
    see also Humean supervenience
  life of, 3–13
    in Australia, 3, 11–13
    in correspondence, 206–208
    early life, 3–8
    education, 8–11, 328–330
    hobbies, 535–536
    parents, 3–4
    work with linguists, 329–330, 339–341
  methodology of
    common sense, 27–29, 231–232, 296–297
  positions of *see* composition, unrestricted; functionalism, analytic; Humean supervenience; materialism; modality, realism; nominalism, about properties; persistence, perdurantism; realism, scientific
Lewis, John and Ewart, 4, 8
Lewis, Stephanie, 35, 42–58
Liao, Shen-Yi, 305
lies, 325
linguistics, 339–341
  see also semantics
location problem, 109
  see also Jackson, Frank
Locke, John, 66, 69–70
Loewer, Barry, 120, 191–192, 273–274, 285–286, 291, 378

logic
higher-order, 466–467
modal, 18–20, 221
plural *see* plurality
see also decision theory; mathematics; mereology; set theory
quantified modal logic (QML), 221–222
see also counterfactual(s); counterpart theory; modality; semantics, possible worlds
Lycan, William, 350, 526

MacBride, Fraser, 184n, 235n, 255
MacFarlane, John, 133n
magnetism, reference *see* semantic theories, eligibility
magnitude(s), 112–114, 264–265, 272
and chance, 280–284, 288–291
and Humean supervenience, 112–114, 284, 291
intrinsic vs extrinsic, 112–114
and laws of nature, 264–265, 272
Malebranche, Nicholas, 70
manifold, 45
see also substantivalism, manifold
Markerese, 331
see also semantics, truth-conditional
Marmor, Andrei, 325
Martin, Charles, 159–160, 307
Mary's room *see* knowledge argument
materialism, 26, 520–522, 524–526
see also Lewis, David, conceptual analysis, naturalism
mathematics
model theory, 19
structuralism about, 163, 386–389, 390–392, 462–463
see also logic; set theory
Maudlin, Tim, 45, 108, 191
Maynard Smith, John, 319
McCawley, James, 329, 330, 332–333, 335, 340
McDermott, Michael, 300, 302
McGee, Vann, 432
McHarry, John, 216–218
McKenzie, Kerry, 26
McLaughlin, Brian, 151–152, 260
Meacham, Christopher J. G., 280
meaning
as conventional, 320–326, 334
improvised, 321–322
structured, 332–333
see also semantics
Melia, Joseph, 135n, 151, 153
Menzies, Peter, 107
mereology, 383–386, 460–461, 463–464
formal, 253–254
mereological essentialism, 15–16, 163, 176, 254–255
and set theory, 252, 382–395, 459–469
simples, extended, 385
see also composition
Merricks, Trenton, 257
metaphysics, 26–28, 40–42, 55–58, 112
Mill, John Stuart, 537, 549–550
Miller, Kristie, 43
Millikan, Ruth, 324
mind–body identity *see* identity theory of mind
minor entity(ies) *see* hole(s)
modality, 71–74, 194
combinatorialism, 143–154, 171, 420–421
compossibility, 19
contextualism about *de re*, 173, 225–227
de dicto, 19, 220
de re, 19, 161, 221–230
epistemic, 197–198
ersatzism, 19, 72, 81, 392–395
fictionalism, 18
practical, 197–198
primitivism, 18, 82
realism, 20, 80–82, 89, 200–201, 216–218, 393, 500–501
see also counterpart relation; counterpart theory; Humean supervenience
Montague, Richard, 11, 88, 329–330, 333
Monton, Bradley, 217
mood, 333
Moore, George Edward, 27
Moore, Michael, 309
Moorean facts *see* Lewis, David, methodology of, common sense
morality *see* value
Morgenstern, Oskar, 316
mosaic, Humean *see* Humean supervenience
multiple realizability *see* functionalism
mutually assured destruction (MAD), 554–555

Nash, John, 317
Nash equilibrium *see* game theory, Nash equilibrium
necessary connection(s), 17
see also Hume's dictum (HD)
necessity
a posteriori, 89, 507–508
see also modality
Newcomb's problem, 448–450
see also decision theory
Newmeyer, Frederick, 329
Ney, Alyssa, 153
Nola, Robert, 38
Nolan, Daniel, 26, 28, 34, 38n, 560n, 561n
nominalism, 35, 61–62, 65–66, 461–462
about properties, 62
about sets, 163
see also realism
Norton, John, 45
noun phrases, 336–337
Nozick, Robert, 216, 448

obligation, moral, 552, 558–559
Ockham's Razor, 61–62
Ockham, William 61–62, 71
Oddie, Graham, 442
O'Leary-Hawthorne, John *see* Hawthorne, John
omission *see* causation, by omission
Onea, Edgar, 350

ontological commitment and innocence, 165, 459–461, 463
see also composition, ontological innocence
Open Question argument, 546
Oresme, Nicole, 67–68

pain, 33, 85–87, 279, 504–512, 515–516, 529–530
see also C-fiber
Pargetter, Robert, 111, 284–285
Parsons, Terence, 336
Partee, Barbara, 88, 329, 334, 336, 339, 341, 345
parthood, 172, 250–251, 253–254, 469
see also composition; mereology
Paul, Laurie A., 153, 157, 227, 231
Pearl, Judea, 301–302
penal substitution, 213–215, 555
performative(s), 346
permutation argument, 369, 371, 377
see also semantic theories, eligibility
Perrault, C. Raymond, 358
Perry, John, 88–90
persistence, 65–68, 237–248, 254–256
endurantism, 238–244, 255–256
and Humean supervenience, 238–242
perdurantism, 84, 237–248, 254–255
stage theory, 244–248
temporary intrinsics, 242–244
person-stages, 241
Pesetsky, David, 323
Pettit, Philip, 440, 510–511
phenomenal character, 514–516
physicalism *see* materialism
physics
Newtonian, 262–263
quantum, 27, 108, 278
see also Best System Account (BSA); Lewis, David, conceptual analysis, science
Pietroski, Paul, 334
Plantinga, Alvin, 210–212
Plato, 62–63, 68–70
pluralism, value *see* value, pluralism
plurality, 390, 392, 407–408, 493
Portner, Paul, 341
possessive(s), 349–350
possibility *see* modality
Postal, Paul, 331–332
Potts, Chris, 350, 353
pragmatics *see* implicature, conversational; presupposition
preference(s) *see* decision theory
presentism, 243
presupposition, 347–363
accommodation, 348–362
classical theory of, 347–349
failure, 359–362
projection, 347–348
trigger, 349–353
Price, George, 319
Price, Huw, 196, 198, 202, 356, 440
Principal Principle (PP), 111–112, 194–195, 282–284
Prior, Arthur, 72
probability(-ies), conditional, 425–428
prominence, prosodic *see* focus
properties, 60–64, 389–392
alien, 102, 128, 145, 521
dispositionalism, 420–421
elite *see* properties, natural
intrinsic, 178, 242–244, 491–497
natural, 103–104, 116, 140, 146, 368, 373–374, 390–392, 420–422, 520
relative naturalness, 370, 373, 374–375
surrogate, 126–130
see also universal(s)
propositions, 161
structured, 399
unstructured, 400
prosody *see* focus
Pryor, Jim, 104
psychology, folk, 27, 434, 504–505, 507–514, 523–529
see also functionalism, analytic
Putnam, Hilary, 10, 334, 368, 514–516
Putnam's paradox *see* permutation argument

quality(-ies), primary and secondary, 63
quantifier(s), 336–337
A- and D-quantifiers, 483
adverbial *see* adverbs of quantification
determiner, quantificational *see* determiner
generalized, 336–337
see also domain restriction
qua objects, 15–16, 173–174, 229–230
see also counterpart theory
question(s), 352, 357
question under discussion (QUD), 357–359, 477, 483–484
see also scoreboard
Quine, Willard Van Orman, 20, 83, 320, 382–383, 391
on modality, 161, 221, 224–225
Quinn, Phil, 207

Ramseyan humility, 495–497
Ramsey sentences *see* Lewis, David, conceptual analysis, Ramseyfication
Ramsey test, 428
see also Adams's Thesis; probability(-ies), conditional
realism
mathematical, 20
metaphysical, 19–20, 60–62
modal *see* modality, realism
normative, 543–545
scientific, 83–84
see also anti-realism; nominalism
recombination, free *see* Hume's dictum (HD); modality, combinatorialism
reduction, 419–420
reference
indeterminacy of *see* indeterminacy, referential
see also semantic theories
reference magnetism *see* semantic theories, eligibility
relationism, 44–45, 47–50
relativism, moral, 543–545

relativity, general, 45, 53–54
representation, 197–199
Rescorla, Michael, 315
Restall, Greg, 183n
Roberts, Craige, 349–350, 353, 357
Roberts, John, 119–120
Robinson, Denis, 113
Rooth, Mats, 336, 352
Rosenbaum, Peter, 329
Rosen, Gideon, 15, 135n
Ross, Haj, 330
Ross, John, 333–334
Russell, Bertrand, 83, 88, 320
Ryle, Gilbert, 504

salience, 118, 120, 125–126
Salmon, Nathan, 90
Savage, Leonard, 289–290, 513
Schaffer, Jonathan, 85, 143, 147, 284–285, 302, 373, 374, 478, 482–484, 485, 511
Schelling, Thomas, 11, 320
Schiffer, Stephen, 334
Schwarz, Wolfgang, 103, 528–529, 531n
scoreboard, 346, 356–359, 484
Scott, Dana, 19–21
Scotus, John Duns, 74
semantics
  dynamic, 331, 337
  Generative vs Interpretive, 333
  possible worlds, 18, 41
  and pragmatics, 321–322, 345
  and psychology, 334
  truth-conditional, 331–334
semantic theories, 129–130, 367–379
  eligibility, 368–379
  externalism, 514–515
  fit, 367–368, 377–378
set theory, 16, 18–20, 252, 383–395, 460–463, 464–469
  constructionalism, 382–384, 389, 392–395
  essentialism, 15
  singletons, 258–259, 383–385, 386–389
  see also axiom(s), of set theory; mereology, and set theory
Shoemaker, Sydney, 141, 420, 497
Si, Aung, 319
Sider, Ted, 101, 144, 152–153, 370
signal, 318–319
  costly, 325
similarity, 222–223
  between individuals *see* properties, natural
  between worlds, 20–21, 84–85, 142–143, 416
  see also counterfactual(s); counterpart relation
Simons, Mandy, 349–350, 353, 357
simplicity, 265, 371–372
  as user-friendliness, 273–274
  see also Best System Account
Skyrms, Brian, 316
Sleeping Beauty, 474
Slote, Michael, 223
Smart, Jack, 10
Smith, Michael, 34
Sobel, Jordan Howard, 450
Sosa, Ernest, 231–232, 476
space *see* relationism; substantivalism
spatiotemporalism, 101–102
speech act(s) *see* performative(s)
Spence, Michael, 325
Stalnaker, Robert, 37, 84, 88, 142, 335, 338, 339, 346, 347–348, 357–358, 416–417, 426–428, 431–433, 509, 514, 528
Stalnaker's Thesis *see* probability(-ies), conditional
Stanley, Jason, 352, 475, 484, 526
states of affairs, 15–17, 228–229
  atomic, 160, 164
  Lewis's critique of, 15–17, 169–172
  as thick particulars, 165
  and thing ontology, 165–166
  totality, 166–167
  see also Truthmaker Principle; truthmaker(s)
Steedman, Mark, 336
Stenius, Erik, 333
Stokhof, Martin, 354
Stone, Matthew, 324, 330, 347–348
strategy,
  evolutionarily stable, 319
  joint, 317, 319
Strawson, Galen, 105
Strawson, Peter, 8, 328
Strevens, Michael, 284, 289–291
Su, Songkun, 319
substance(s) *see* hylomorphism
substantivalism, 45–46, 50–55
  manifold, 45, 55–56
  metric field, 45–46, 54–55
  tensor field, 46, 50–53
Sugden, Robert, 324–325
supervenience, 102–103, 259–260, 520–521
  see also Humean supervenience
Swoyer, Chris, 141
symmetry, 201, 290–291
  see also asymmetry
synonymy *see* analyticity
syntax *see* grammar; semantics
Szabo, Zoltan, 352
Számadó, Szabolcs, 319

Tautz, Jürgen, 319
Taylor, Barry, 119–120, 380n
Thau, Michael, 282–283
theodicy, 209–213, 216–218
Thomason, Rich, 347–348, 354–355
Thomasson, Annie, 204n
three-dimensionalism *see* persistence, endurantism
time *see* counterpart relation, modal vs counterpart relation temporal; persistence; presentism
toleration, 549–552
Tonhauser, Judith, 349–350, 353, 357
Tooley, Michael, 207
transcendental argument, 498, 500–501
triviality argument, 425–433, 452–455
  see also Adams's Thesis; probability(-ies), conditional

trope(s), 16–17, 165
  non-transferability of, 165
truth
  correspondence theory, 167–169
  redundancy theory, 168
  supervenience on being, 101, 180–183
    see also truthmaker(s)
Truthmaker Principle, 159–183, 196, 228, 230
  as non-fundamental, 181–182
  plural version, 163
  and theories of truth, 167–169
  see also truthmaker(s)
truthmaker(s), 15–17, 159–183, 194–195, 229–230
  entailment principle, 162
  for general propositions, 166
  for necessary truths, 162
  necessitarianism, 161–162
  for negative existentials, 166–167, 170, 171, 173, 175, 228–230
  parthood principle, 163–164
  as states of affairs, 15–17, 164–165
  as things, 175–176
  truthmaker argument, 165
  see also counterpart theory; grounding; states of affairs
Twin Earth, 110, 514–515
two-dimensional semantics, 88–89

Unger, Peter, 558
universal(s), 238
  structural, 251–252
  see also properties
Uzquiano, Gabriel, 255

vagueness, 252–253
  see also composition, unrestricted; indeterminacy
value, 536–547
  expected, 447, 512
  pluralism, 539–541
  qualified attitude account, 536–538, 543, 545–546
  see also decision theory
van der Sandt, Robert, 347–348
Vanderschraaf, Peter, 324–325
van Fraassen, Bas, 32, 83
van Inwagen, Peter, 255–258
van Kuppevelt, Jan, 357
van Lambalgen, Michael, 334
van Rooij, Robert, 357
Varzi, Achille, 42–43, 50–54, 256
verbs, factive, 349–350
  see also presupposition
virtue ethics, 554
virtue of, in *see* grounding
von Fintel, Kai, 347–348
von Frisch, Karl, 319
von Neumann, John, 316, 468
von Stechow, Arnim, 332, 340

Weatherson, Brian, 103, 108–109, 328, 560n
Wiggins, David, 544
will, free *see* freedom
Williams, Bernard, 196
Williams, Donald Cary, 219n
Williams, J. Robert G., 129–130, 253, 373
Williamson, Timothy, 526
Wilson, Deirdre, 350
Wilson, Jessica, 149–150
Winsberg, Eric, 131n
Wittgenstein, Ludwig, 160, 202, 226
worldmate, 499–501
world-state(s), 81, 84

Xue, Jingyang, 350

Yablo, Stephen, 301–302

Zermelo-Frankel set theory with choice (ZFC) *see* axiom(s), of set theory
Zhang, Shaowu, 319
zombies, philosophical 515–516